

VISIT...

LANZAROTE
Caliente.COM

MONOGRAPHS ON NUMERICAL ANALYSIS

General editors

E. T. GOODWIN, L. FOX

THE ALGEBRAIC EIGENVALUE PROBLEM

by

J. H. WILKINSON

CLARENDON PRESS, OXFORD

1965

Дж. Х. УИЛКИНСОН

АЛГЕБРАИЧЕСКАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ

Перевод с английского
В. В. ВОЕВОДИНА и В. Н. ФАДДЕЕВОЙ

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1970

Книга посвящена численным методам решения задач алгебры, в основном методам отыскания собственных значений матриц и соответствующих им собственных векторов. Однако в ней достаточно полно представлены методы решения и других задач алгебры, таких как решение систем линейных алгебраических уравнений, отыскание корней функций и т. д. Книга состоит из девяти глав.

В главах I, II излагается вспомогательный материал, содержащий основы теории линейной алгебры.

Главы III, IV, V содержат анализ ошибок округления простейших операций, некоторых линейных преобразований, методов решения систем линейных уравнений. Материал этих глав является фундаментом всех дальнейших исследований.

Главы VI, VII содержат описание и анализ ошибок методов приведения общей матрицы к матрице специального вида. Рассматриваются методы решения полной проблемы собственных значений этих матриц.

В главах VIII, IX излагается обширный материал с анализом ошибок по решению полной проблемы собственных значений степенными методами.

Таблиц 15, рисунков 4, библиографических ссылок 146.

Дж. Х. Уилкинсон

Алгебраическая проблема собственных значений

М., 1970 г., 564 стр. с илл.

Редакторы: И. М. Овчинникова и Г. С. Росляков

Техн. редактор И. Ш. Аксельрод

Корректор Т. С. Вайсберг

Сдано в набор 11/VII 1969 г. Подписано к печати 30/I 1970 г.
Бумага 70×108¹/₁₆. Физ. печ. л. 35,5. Условн. печ. л. 49,7.
Уч.-изд. л. 44,59. Тираж 6450 экз. Цена книги 3 р. 41 к.
Заказ № 1054.

Издательство «Наука»

Главная редакция физико-математической литературы
Москва, В-71, Ленинский проспект, 15.

Московская типография № 16 Главполиграфпрома
Комитета по печати при Совете Министров СССР.
Москва, Трехпрудный пер., 9

ОГЛАВЛЕНИЕ

Предисловие переводчиков	15
Предисловие автора	16
Глава 1. Теоретическое обоснование	19
Введение	19
Определения	20
Собственные значения и собственные векторы транспонированной матрицы	21
Различные собственные значения	22
Преобразования подобия	23
Кратные собственные значения и канонические формы матриц общего вида	24
Неполная система собственных векторов	26
Жорданова (классическая) каноническая форма	27
Элементарные делители	28
Сопровождающая матрица характеристического полинома	28
Полные матрицы	29
Каноническая (рациональная) форма Фробениуса	31
Связь между жордановой и фробениусовой каноническими формами	32
Преобразования эквивалентности	32
λ -матрицы	33
Элементарные операции	33
Каноническая форма Смита	34
Наибольший общий делитель k -строчных миноров λ -матрицы	36
Инвариантные множители ($A - \lambda I$)	36
Треугольная каноническая форма	37
Эрмитовы и симметричные матрицы	38
Элементарные свойства эрмитовых матриц	39
Комплексные симметричные матрицы	40
Приведение к треугольной форме унитарными преобразованиями	40
Квадратичные формы	40
Необходимые и достаточные условия положительной определенности	41
Дифференциальные уравнения с постоянными коэффициентами	43
Решения, соответствующие нелинейным элементарным делителям	43
Дифференциальные уравнения высшего порядка	45
Уравнения второго порядка специального вида	46
Явное решение уравнения $B\ddot{y} = -Ay$	47
Уравнения вида $(AB - \lambda I)x = 0$	48
Минимальный полином вектора	49
Минимальный полином матрицы	49
Теорема Кэли — Гамильтона	50
Соотношения между минимальным полиномом и каноническими формами	51
Корневые векторы	53
Элементарные преобразования подобия	54
Свойства элементарных матриц	55
Приведение к треугольной канонической форме элементарными преобразованиями подобия	56
Элементарные унитарные преобразования	57
Элементарные унитарные эрмитовы матрицы (матрицы отражения)	58

Приведение к треугольному виду элементарными унитарными преобразованиями	59
Нормальные матрицы	60
Коммутирующие матрицы	61
Собственные значения AB	63
Векторные и матричные нормы	63
Подчиненные матричные нормы	64
Евклидова и спектральная нормы	65
Нормы и пределы	67
Некоторые оценки без использования бесконечных матричных рядов	69
Дополнительные замечания	69
Глава 2. Теория возмущений	70
Введение	70
Теорема Островского о непрерывности собственных значений	71
Алгебраические функции	72
Численные примеры	73
Теория возмущений для простых собственных значений	74
Возмущение соответствующих собственных векторов	74
Матрица с линейными элементарными делителями	75
Возмущения первого порядка собственных значений	75
Возмущения первого порядка собственных векторов	76
Возмущения высоких порядков	77
Кратные собственные значения	77
Теоремы Гершгорина	78
Теория возмущений, основанная на теоремах Гершгорина	79
Возмущения, соответствующие общему распределению нелинейных делителей	85
Теория возмущений для собственных векторов матриц, основанная на исследовании жордановой канонической формы	86
Возмущения собственных векторов, соответствующих кратным собственным значениям (линейные элементарные делители)	88
Ограничения теории возмущений	88
Соотношения между s_i	89
Обусловленность вычислительных проблем	90
Числа обусловленности	90
Спектральное число обусловленности матрицы по отношению к проблеме собственных значений	90
Свойства спектрального числа обусловленности	91
Инвариантные свойства чисел обусловленности	92
Очень плохо обусловленные матрицы	93
Теория возмущений для вещественных симметричных матриц	96
Несимметричные возмущения	96
Симметричные возмущения	96
Классическая техника	97
Симметричная матрица с рангом единица	99
Экстремальные свойства собственных значений	100
Минимаксные свойства собственных значений	101
Собственные значения суммы двух симметричных матриц	103
Практические применения	104
Дальнейшие применения принципа минимакса	104
Теорема разделения	104
Теорема Виландта — Гофмана	105
Дополнительные замечания	109

Глава 3. Анализ ошибок	110
Введение	110
Операции с фиксированной запятой	110
Накопление скалярного произведения	111
Операции с плавающей запятой	112
Упрощенные выражения для границ ошибок	113
Оценки ошибок для некоторых основных вычислений с плавающей запятой	113
Оценки норм матриц ошибок	115
Накопление скалярных произведений в арифметике с плавающей запятой	115
Оценки ошибок для некоторых основных вычислений $fl_2()$ с двойным числом разрядов	116
Вычисление квадратных корней	118
Блочнo-плавающие векторы и матрицы	118
Основные ограничения t -разрядных вычислений	119
Методы определения собственных значений, основанные на подобных преобразованиях	121
Анализ ошибок методов, основанных на элементарных неунитарных преобразованиях	123
Анализ ошибок методов, основанных на элементарных унитарных преобразованиях	124
Преимущество унитарных преобразований	125
Вещественные симметричные матрицы	126
Ограничения унитарных преобразований	127
Анализ ошибок вычисления плоских вращений с плавающей запятой	128
Умножение на плоское вращение	130
Умножение на последовательность плоских вращений	131
Ошибка в произведении приближенных плоских вращений	135
Ошибки в подобных преобразованиях	136
Симметричные матрицы	137
Плоские вращения в арифметике с фиксированной запятой	139
Другое вычисление $\sin \theta$ и $\cos \theta$	140
Левое умножение на приближенное вращение с фиксированной запятой	141
Умножение на последовательность плоских вращений	142
Вычисленное произведение приближенной последовательности плоских вращений	143
Ошибки в подобных преобразованиях	144
Общее замечание об оценках ошибки	146
Матрицы отражения с плавающей запятой	147
Анализ ошибок вычисления матрицы отражения	148
Численный пример	151
Левое умножение на приближенную матрицу отражения	152
Умножение на последовательность приближенных матриц отражения	154
Неунитарные элементарные матрицы, аналогичные плоским вращениям	156
Неунитарные элементарные матрицы, аналогичные матрицам отражения	157
Левые умножения на последовательность неунитарных матриц	159
Априорные оценки ошибок	160
Отклонение от нормальности	160
Простые примеры	162
Апостериорные оценки	163
Апостериорные оценки для нормальных матриц	163
Отношение Релея	165
Ошибка в отношении Релея	166
Эрмитовы матрицы	167

Патологически близкие собственные значения	169
Ненормальные матрицы	171
Анализ ошибок для полной собственной системы	172
Численный пример	173
Условия, ограничивающие достижимую точность	173
Нелинейные элементарные делители	174
Приближенные инвариантные подпространства	176
Почти нормальные матрицы	178
Дополнительные замечания	178
Глава 4. Решение линейных алгебраических уравнений	179
Введение	179
Теория возмущений	179
Числа обусловленности	180
Уравновешенные матрицы	181
Простые практические примеры	182
Обусловленность матрицы собственных векторов	183
Явное решение	184
Общие замечания об обусловленности матриц	184
Связь плохой обусловленности и почти вырожденности	185
Ограничения, накладываемые t -разрядной арифметикой	186
Алгоритмы решения линейных уравнений	187
Метод Гаусса	188
Треугольное разложение	189
Структура матриц треугольного разложения	189
Явные выражения для элементов треугольных матриц	190
Несостоятельность метода Гаусса	192
Численная устойчивость	192
Значение перестановок	193
Численный пример	194
Анализ ошибок метода Гаусса	196
Верхние оценки матриц возмущения для арифметики с фиксированной запятой	197
Верхняя оценка для элементов преобразованных матриц	198
Выбор главного элемента по всей матрице	199
Практический процесс с выбором главного элемента по столбцу	200
Анализ ошибок с плавающей запятой	200
Разложение с плавающей запятой без выбора главного элемента	201
Потеря верных знаков	202
Общераспространенная ошибка	203
Матрицы специального вида	203
Метод Гаусса на быстродействующей вычислительной машине	205
Решения, соответствующие различным правым частям	206
Прямое треугольное разложение	206
Связь метода Гаусса с прямым треугольным разложением	207
Примеры несостоятельности и неединственности разложения	208
Треугольное разложение с перестановкой строк	209
Анализ ошибок треугольного разложения	211
Вычисление определителей	212
Разложение Холецкого	213
Произвольные симметричные матрицы	213
Анализ ошибок разложения Холецкого в арифметике с фиксированной запятой	214

Плохо обусловленная матрица	216
Триангуляризация, использующая матрицы отражения	218
Анализ ошибок триангуляризации Хаусхолдера	219
Триангуляризация элементарными устойчивыми матрицами типа M'_{ji}	219
Вычисление главных миноров	220
Триангуляризация плоскими вращениями	221
Анализ ошибок преобразования Гивенса	222
Единственность ортогональной триангуляризации	223
Ортогонализация Шмидта	224
Сравнение методов триангуляризации	226
Обратная подстановка	228
Высокая точность вычисленного решения треугольной системы уравнений	230
Решение общей системы уравнений	232
Вычисление общей обратной матрицы	233
Точность вычисленных решений	233
Плохо обусловленные матрицы, которые дают немалые ведущие элементы	234
Итерационное улучшение приближенного решения	235
Влияние ошибок округления на итерационный процесс	236
Итерационный процесс в вычислении с фиксированной запятой	236
Простой пример итерационного процесса	237
Общие замечания по итерационному процессу	239
Родственные итерационные процессы	240
Ограниченность итерационного процесса	240
Строгое обоснование итерационного метода	241
Дополнительные замечания	241
Глава 5. Эрмитовы матрицы	243
Введение	243
Классический метод Якоби для вещественных симметричных матриц	243
Быстрота сходимости	244
Сходимость к фиксированной диагональной матрице	245
Циклический метод Якоби	246
Круги Гершгорина	247
Асимптотическая квадратичная сходимость методов Якоби	247
Ближие и кратные собственные значения	248
Численные примеры	250
Вычисление $\cos \theta$ и $\sin \theta$	251
Более простое определение углов вращения	253
Циклический метод Якоби с преградами	254
Вычисление собственных векторов	254
Численный пример	255
Анализ ошибок метода Якоби	255
Точность вычисленных собственных векторов	256
Оценки ошибок для вычислений с фиксированной запятой	257
Организационные проблемы	257
Метод Гивенса	258
Процесс Гивенса на вычислительной машине с двухступенчатой памятью	259
Анализ ошибок процесса Гивенса в арифметике с плавающей запятой	261
Анализ ошибок в арифметике с фиксированной запятой	262
Численный пример	262
Метод Хаусхолдера	264
Использование симметрии	266
Некоторые соображения о необходимой величине памяти	267

Процесс Хаусхолдера на вычислительной машине с двухступенчатой памятью	267
Метод Хаусхолдера в арифметике с фиксированной запятой	268
Численный пример	269
Анализ ошибок метода Хаусхолдера	270
Собственные значения симметричной трехдиагональной матрицы	272
Свойство последовательности Штурма	272
Метод деления пополам	274
Численная устойчивость метода деления пополам	275
Численный пример	277
Общие замечания о методе деления пополам	277
Малые собственные значения	278
Ближкие собственные значения и малые β_i	279
Вычисление собственных значений в арифметике с фиксированной запятой	283
Вычисление собственных векторов трехдиагональной матрицы	285
Неустойчивость явного выражения для собственного вектора	286
Численные примеры	288
Обратная итерация	290
Выбор начального вектора b	291
Анализ ошибок	292
Численный пример	294
Ближкие собственные значения и малые β_i	295
Линейно независимые векторы, соответствующие совпадающим собственным значениям	296
Другой метод вычисления собственных векторов	298
Численный пример	299
Замечания о проблеме собственных значений для трехдиагональных матриц	300
Завершение методов Гивенса и Хаусхолдера	300
Сравнение методов	302
Квазисимметричные трехдиагональные матрицы	302
Вычисление собственных векторов	303
Уравнения вида $Ax = \lambda Bx$ и $ABx = \lambda x$	304
Численный пример	305
Одновременное приведение A и B к диагональному виду	307
Трехдиагональные A и B	307
Комплексные эрмитовы матрицы	308
Дополнительные замечания	310
Глава 6. Приведение общих матриц к матрицам специального вида	311
Введение	311
Метод Гивенса	311
Метод Хаусхолдера	313
Соображения о необходимой величине памяти	315
Анализ ошибок	315
Связь методов Гивенса и Хаусхолдера	316
Элементарные устойчивые преобразования	318
Значение перестановок	319
Прямое приведение к форме Хессенберга	321
Введение перестановок	322
Численный пример	323
Анализ ошибок	327
Дальнейший анализ ошибок	329

Плохая определенность матрицы Хессенберга	331
Приведение к форме Хессенберга матрицами типа M'_{ji}	331
Метод Крылова	332
Метод Гаусса с исключением по столбцам	332
Практические трудности	333
Обусловленность C для некоторых стандартных расположений собственных значений	334
Начальный вектор высоты, меньшей чем n	336
Практическое применение	337
Обобщенные процессы Хессенберга	338
Срыв обобщенного процесса Хессенберга	339
Метод Хессенберга	340
Практическая процедура	341
Связь метода Хессенберга и предыдущих методов	341
Метод Арнольди	342
Практические рассуждения	343
Значение переортогонализации	345
Метод Ланцоша	347
Срыв процесса	348
Численный пример	349
Практический процесс Ланцоша	349
Численный пример	350
Общие замечания к несимметричному процессу Ланцоша	352
Симметричный процесс Ланцоша	352
Приведение матриц Хессенберга к более компактному виду	353
Приведение нижней матрицы Хессенберга к трехдиагональному виду	353
Использование перестановок	354
Влияние малого ведущего элемента	355
Анализ ошибок	356
Процесс Хессенберга в применении к нижней матрице Хессенберга	358
Связь процесса Хессенберга и процесса Ланцоша	358
Приведение матрицы общего вида к трехдиагональному виду	359
Сравнение с методом Ланцоша	360
Повторное исследование приведения к трехдиагональному виду	360
Приведение матриц в верхней форме Хессенберга к форме Фробениуса	361
Влияние малых ведущих элементов	362
Численный пример	363
Общие замечания об устойчивости	363
Специальная верхняя форма Хессенберга	364
Прямое определение характеристического полинома	365
Дополнительные замечания	365
Глава 7. Собственные значения матриц специального вида	367
Введение	367
Явно заданная полиномиальная форма	367
Числа обусловленности явно заданных полиномов	369
Несколько типичных расположений корней	370
Окончательная оценка метода Крылова	374
Общие замечания о явно заданных полиномах	374
Трехдиагональные матрицы	376
Определители матриц Хессенберга	378
Влияние ошибок округления	379
Накопление с плавающей запятой	380

Вычисление при помощи ортогональных преобразований	381
Вычисление определителей матриц общего вида	382
Обобщенная проблема собственных значений	383
Косвенное определение характеристического полинома	383
Метод Лаврентье	384
Итерационные методы, основанные на интерполяции	385
Асимптотическая скорость сходимости	386
Кратные корни	388
Обращение функционального соотношения	389
Метод деления отрезка пополам	390
Метод Ньютона	391
Сравнение метода Ньютона с методом интерполяции	392
Методы, дающие кубическую сходимость	393
Метод Лагерра	393
Комплексные корни	395
Комплексно сопряженные корни	397
Метод Берстоу	398
Обобщенный метод Берстоу	399
Практические рассуждения	400
Влияние ошибок округления на асимптотическую сходимость	401
Метод деления отрезка пополам	401
Последовательная линейная интерполяция	403
Кратные и патологически близкие собственные значения	404
Другие интерполяционные методы	405
Методы, в которых используются производные	407
Критерий достижения корня	408
Влияние ошибок округления	408
Удаление вычисленных корней	410
Исчерпывание для матриц Хессенберга	410
Исчерпывание для трехдиагональной матрицы	412
Исчерпывание при помощи вращений или устойчивых элементарных преобразований	414
Устойчивость исчерпывания	415
Общие замечания об исчерпывании	417
Удаление вычисленных нулей	417
Удаление вычисленного квадратичного множителя	418
Общие комментарии о методах удаления	418
Асимптотическая скорость сходимости	419
Сходимость в целом	420
Комплексные корни	422
Рекомендации	423
Комплексные матрицы	424
Матрицы, содержащие независимый параметр	424
Дополнительные замечания	425
Глава 8. LR- и QR-алгоритмы	426
Введение	426
Вещественные матрицы с комплексными собственными значениями	427
LR-алгоритм	428
Доказательство сходимости последовательности A_s	429
Положительно определенные эрмитовы матрицы	433
Комплексно сопряженные собственные значения	434
Введение перестановок	437

Численный пример*	438
Сходимость модифицированного процесса	439
Предварительная подготовка исходной матрицы	439
Инвариантность верхней формы Хессенберга	440
Одновременное действие со строками и столбцами	441
Ускорение сходимости	442
Объединение сдвигов	443
Выбор сдвига	444
Исчерпывание матрицы	445
Экспериментальная проверка сходимости	446
Улучшенный выбор сдвига	447
Комплексно сопряженные собственные значения	448
Критика модифицированного LR -алгоритма	449
QR -алгоритм	450
Сходимость QR -алгоритма	451
Формальное доказательство сходимости	452
Нарушение порядка собственных значений	453
Равные по модулю собственные значения	454
Другое доказательство для LR -метода	455
Практическое применение QR -алгоритма	457
Сдвиг	458
Разложение матриц A_s	458
Численный пример	460
Практическая процедура	460
Устранение комплексно сопряженных сдвигов	461
Двойной QR -шаг с использованием матриц отражения	464
Вычислительные детали	465
Разложение матриц A_s	466
Прием двойного сдвига для LR	467
Сравнение LR - и QR -алгоритмов	468
Кратные собственные значения	469
Специальное использование процесса исчерпывания	472
Симметричные матрицы	472
Связь между LR - и QR -алгоритмами	473
Сходимость LR -алгоритма Холецкого	474
Кубическая сходимость QR -алгоритма	476
Сдвиг в LR -алгоритме Холецкого	477
Срыв факторизации Холецкого	478
Кубически сходящийся LR -процесс	479
Ленточные матрицы	481
QR -разложение ленточной матрицы	483
Анализ ошибок	486
Несимметричные ленточные матрицы	487
Одновременное разложение и перемножение в QR -алгоритме	489
Уменьшение ширины ленточной матрицы	491
Дополнительные замечания	491
Глава 9. Итерационные методы	493
Введение	493
Степенной метод	493
Прямые итерации одного вектора	494
Сдвиг начала	495
Влияние ошибок округления	496

Изменение p	498
Специальный выбор p	499
Процесс ускорения Эйткена	500
Комплексно сопряженные собственные значения	501
Вычисление комплексного собственного вектора	502
Сдвиг начала	503
Нелинейные делители	504
Одновременное определение нескольких собственных значений	504
Комплексные матрицы	505
Исчерпывание	505
Исчерпывание, основанное на преобразованиях подобия	506
Исчерпывание, использующее инвариантные подпространства	507
Исчерпывание при помощи устойчивых элементарных преобразований	508
Исчерпывание при помощи унитарных преобразований	509
Численная устойчивость	510
Численный пример	512
Устойчивость унитарных преобразований	513
Исчерпывание при помощи неподобных преобразований	514
Общее приведение, использующее инвариантные подпространства	517
Практическое приложение	518
Ступенчатые итерации	520
Точное определение комплексно сопряженных собственных значений	522
Очень близкие собственные значения	523
Метод ортогонализации	523
Аналог ступенчатых итераций, использующий ортогонализацию	524
Биортогонализация	525
Численный пример	526
Процесс уточнения Рундсона	529
Возведение матрицы в степень	530
Численная стабильность	531
Использование полиномов Чебышева	532
Общая оценка методов, основанных на прямых итерациях	533
Обратные итерации	534
Анализ ошибок при обратных итерациях	534
Общие комментарии к анализу	536
Дальнейшее уточнение собственных векторов	536
Нелинейные элементарные делители	539
Обратные итерации с матрицами Хессенберга	540
Вырожденные случаи	541
Обратные итерации с ленточными матрицами	541
Комплексно сопряженные собственные векторы	542
Анализ ошибок	544
Численный пример	545
Обобщенная проблема собственных значений	547
Изменение приближенных собственных значений	547
Уточнение системы собственных значений	549
Численный пример	551
Уточнение собственных векторов	552
Комплексно сопряженные собственные значения	554
Совпадающие и патологически близкие собственные значения	555
Замечания о программах на АСЕ	557
Дополнительные замечания	558
Литература	559

ПРЕДИСЛОВИЕ ПЕРЕВОДЧИКОВ

Монография Дж. Х. Уилкинсона «Алгебраическая проблема собственных значений» посвящена в основном численным методам отыскания собственных значений матриц и соответствующих им собственных векторов. Однако в ней достаточно полно представлены методы решения и других задач алгебры, таких, как решение систем линейных алгебраических уравнений, отыскание корней функций и т. д.

Решению задач алгебры уделяется много внимания в существующей литературе, но эта книга выгодно отличается от большинства современных пособий. Впервые изложение всех методов сопровождается детальным анализом влияния ошибок округления результатов промежуточных вычислений на точность решения рассматриваемых задач. Наличие такого анализа делает данную книгу незаменимым помощником в практической работе. В книге приводится много весьма интересных примеров, характеризующих различные патологические явления в численных методах, что позволяет глубже понять природу самих методов.

Главы 1, 2, 6—9 переведены В. Н. Фаддеевой, главы 3—5 переведены В. В. Воеводиным.

ПРЕДИСЛОВИЕ АВТОРА

Решение алгебраической проблемы собственных значений долго привлекало мое внимание, потому что оно очень хорошо выявляет разницу между тем, что может быть названо классической математикой и практическим численным анализом. Проблема собственных значений имеет обманчиво простую формулировку, и теоретическое ее обоснование известно уже в течение многих лет; в то же время определение точных решений представляет обширное многообразие нерешенных проблем.

Предложение написать книгу на эту тему в серии монографий по численному анализу было сделано мне профессором Фоксом и доктором Гудвином после моих ранних опытов на автоматических цифровых машинах. Возможно, что ее написание осталось бы благим намерением, если бы не приглашение профессора Гивенса принять участие в симпозиуме в Детройте в 1957 г., которое привело к последующему приглашению прочесть лекции на тему «Практическая техника для решения линейных систем и вычисления собственных значений и собственных векторов» в летней школе, организованной Мичиганским университетом. Для этих лекций необходимо было ежегодно готовить материал, что оказалось очень ценным, так как, таким образом, многие результаты этой книги были рассказаны в процессе их получения.

Мой первоначальный план состоял в изложении ряда известных приемов для решения проблемы вместе с критическим обсуждением их достоинств, сопровождаемым, где это возможно, соответствующим анализом ошибок; в 1961 г. рукопись, написанная по этому плану, была готова. Однако в период приготовления рукописи произошел существенный прогресс как в решении проблемы собственных значений, так и в анализе ошибок, и меня все более не удовлетворяли ранние главы. В 1962 г. я принял решение перепланировать всю книгу, изменив первоначальную цель. Я почувствовал, что не имеет смысла рассматривать почти все известные методы и приводить для них анализ ошибок, и решил включить в книгу главным образом те методы, по которым я имел обширный практический опыт. Я включил также дополнительную главу, дающую достаточно общий анализ ошибок, который приложим почти ко всем методам, данным впоследствии. Многолетний опыт убедил меня, что не легко дать надежную оценку метода без его использования и что сравнительно часто незначительное изменение в деталях практического процесса имеет непропорциональное влияние на его эффективность.

Автор книги по численному анализу находится перед трудной проблемой определения круга читателей, которому она должна быть предназначена. Практическая трактовка проблемы собственных значений потенциально интересна широким кругам, в число которых входят инженеры, физики-теоретики, классические математики-прикладники и вычислители, желающие проводить исследования в теории матриц. Книга, адресованная главным образом последним читателям, может оказаться недоступной

для первых. Я ничего не опустил из-за боязни, что это может быть трудным некоторым читателям, но старался описать все настолько элементарно, насколько это позволял предмет. Дилемма была особенно остра в первой главе. Я надеюсь, что данное в ней элементарное изложение не обидит серьезного математика и что он рассмотрит это как грубую наброску классического материала, с которым он должен быть знаком, если он хочет извлечь пользу из книги от начала до конца. Я предположил с самого начала, что читатель знаком с основными понятиями векторного пространства, линейной зависимости и ранга. Великолепное введение к материалу этой книги дано в книге Фокса (L. Fox, *An introduction to numerical linear algebra*, Oxford, 1964.) Научные сотрудники, работающие в этой области, найдут неоценимый источник информации в книге Хаусхолдера (A. S. Householder, *The theory of matrices in numerical analysis*, Blaisdell, 1964).

Мое решение излагать только те методы, с которыми я имел дело, неизбежно привело к тому, что были опущены многие важные алгоритмы. Однако этот пробел менее серьезен, чем он мог бы быть, так как монографии Е. Durand, *Solutions numériques des equations algebriques* (Masson, 1960) и «Вычислительные методы линейной алгебры» Д. К. Фаддеева и В. Н. Фаддеевой (Москва, 1963) содержат очень обширное описание. Однако два пропуска требуют специальной оговорки. Первый—это метод типа Якоби, развитый Эберляйн, и разные его модификации, предложенные независимо Рутисхаузером. Эти методы мне представляются многообещающими, и они вполне могут оказаться наилучшими для решения общей проблемы собственных значений. Я не хотел включить их без предварительного детального изучения, которого они заслуживают. Мое решение было подкреплено тем, что они не полностью покрываются данной мной общей теорией ошибок. Второй пропуск—это общая трактовка QR -алгоритма Рутисхаузера. Он имеет очень широкие приложения, и я почувствовал, что не смогу полностью отдать должное этой работе в пределах, ограниченных областью задачи на собственные значения. Читатель отсылается к соответствующим работам Рутисхаузера и Хенричи. Литература по проблеме собственных значений очень обширна и я составил библиографию главным образом из тех работ, на которые в тексте сделаны явные ссылки. Очень детальная библиография приведена в вышеупомянутых книгах Фаддеева и Фаддеевой и Хаусхолдера.

В процессе изменения плана книги я искушался желанием приложить запись алгоритмов на АЛГОЛе, но решил, что в настоящее время трудности получения верных во всех деталях процедур непреодолимы. Поэтому я употреблял язык классической математики, но принимал форму слов, которую легко перевести в АЛГОЛ.

Материал этой книги получался из многих источников, но я хотел бы особенно отметить работу с моими коллегами по летней школе Мичиганского университета, особенно с Бауэром, Форсайтом, Гивенсом, Хенричи, Хаусхолдером, Ольгой Таусски, Джоном Тодд и Варга. Кроме этой работы, моим главным источником вдохновения был алгоритмический талант Рутисхаузера.

Мне приятно отметить помощь, которую я получил в этом деле от многих друзей. Я рад поблагодарить профессоров Карра и Бартельса за ряд приглашений для прочтения лекций в летней школе в Энн Арборе, которые явились для меня ежегодным стимулом пополнения рукописи. Я особенно обязан профессорам Форсайту и Гивенсу, великодушно обеспечившим возможность расширения двух главных замыслов рукописи. Это

позволило мне извлечь пользу из мнений более широкого круга читателей, чем это было бы возможно. Много ценных советов получено от профессора Форсайта, доктора Варга и мистера Албасини. Полностью рукопись была прочитана доктором Мартином и профессором Фоксом, а первая корректура — доктором Гудвином. Их замечания и советы спасли меня от многих грубых ошибок и от погрешностей стиля.

В трудном деле чтения корректур мне любезно помогла миссис Гудд.

Наконец, я приношу искреннюю благодарность моей жене, отпечатавшей многие варианты рукописи. Без ее терпения и упорства проект был бы, конечно, неосуществлен.

Дж. Х. Уилкинсон

ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ

Введение

1. В этой вводной главе мы дадим краткий обзор классической теории канонических форм и некоторых других теоретических вопросов, которые понадобятся при дальнейшем изложении.

Так как полное изложение выходит за рамки этой главы, мы сконцентрировали наше внимание на том, чтобы показать, как система собственных значений матрицы связана с ее различными каноническими формами. Поэтому изложение несколько отлично от стандартного. Не сделано попыток приводить строгие доказательства всех фундаментальных теорем; вместо этого мы довольствуемся формулировкой результатов, и даем доказательства лишь тогда, когда они особенно тесно связаны с техникой вычислений, используемой впоследствии.

Чтобы избежать повторения объяснений, введем здесь некоторые обозначения, которых будем придерживаться во всей книге.

Будем обозначать матрицы заглавными буквами и, если не оговорено противное, предполагать матрицы квадратными порядка n . Элемент матрицы A , стоящий в позиции (i, j) , будем обозначать через a_{ij} . Векторы будем обозначать маленькими строчными буквами и обычно предполагать их n -мерными; через x_i будем обозначать i -ю компоненту вектора x . Будем обозначать единичную матрицу через I , полагая ее (i, j) -й элемент равным δ_{ij} , символу Кронекера (а не i_{ij}), так что

$$\delta_{ij} = 0 \quad (i \neq j), \quad \delta_{ii} = 1. \quad (1.1)$$

Через e_i будет обозначаться i -й столбец I , а e будет обозначать вектор $e_1 + e_2 + \dots + e_n$, у которого все компоненты равны единице.

Мы часто будем иметь дело с системой n векторов $x_1, x_2, \dots, x_n$. В таком случае x_{ji} будет обозначать j -й элемент x_i , а X — матрицу, имеющую x_i своим i -м столбцом.

Обозначение $|A|$ будет сохранено для матрицы, у которой (i, j) -й элемент равен $|a_{ij}|$, и неравенство

$$|A| \leq |B| \quad (1.2)$$

будет обозначать, что

$$|a_{ij}| \leq |b_{ij}| \quad \text{для всех } i, j. \quad (1.3)$$

В силу того, что $\|A\|$ используется для обозначения нормы A (см. § 52), определитель A будет обозначаться через $\det(A)$. Аналогично диагональная матрица n -го порядка, у которой (i, i) -й элемент равен λ_i , будет обозначаться через $\text{diag}(\lambda_i)$.

Будем обозначать транспонированную матрицу через A^T , так что ее (i, j) -й элемент равен a_{ji} . Соответственно x^T будет обозначать вектор-строку, у которой i -я компонента равна x_i . Так как печатать вектор-строку более

удобно, чем вектор-столбец, мы часто будем вводить вектор-столбец x и затем определять его, задавая явное выражение для x^T .

Мы будем предполагать всюду, что работаем в поле комплексных чисел, хотя в вычислительных процедурах выгоднее оставаться в поле вещественных чисел, там, где это возможно.

Определения

2. Фундаментальная алгебраическая проблема собственных значений состоит в определении тех значений λ , при которых система n однородных линейных уравнений с n неизвестными

$$Ax = \lambda x \quad (2.1)$$

имеет нетривиальное решение. Уравнение (2.1) может быть записано в форме

$$(A - \lambda I)x = 0 \quad (2.2)$$

и при произвольном λ эта система уравнений имеет только решение $x = 0$. Из общей теории совместных линейных алгебраических уравнений известно, что нетривиальное решение существует тогда, и только тогда, когда матрица $(A - \lambda I)$ особенная, т. е.

$$\det(A - \lambda I) = 0. \quad (2.3)$$

Раскрывая определитель в левой части уравнения (2.3) по степеням λ , получим

$$\alpha_0 + \alpha_1 \lambda + \dots + \alpha_{n-1} \lambda^{n-1} + (-1)^n \lambda^n = 0. \quad (2.4)$$

Уравнение (2.4) называется *характеристическим уравнением* матрицы A , а полином в левой части этого уравнения называется *характеристическим полиномом*. Так как коэффициент при λ^n не нуль, и мы предполагаем, что рассматривается поле комплексных чисел, это уравнение всегда имеет n корней. Вообще говоря, корни могут быть комплексными, даже если матрица A вещественная, и корни могут быть любой кратности вплоть до n . Эти корни называются *собственными значениями* или *характеристическими числами* матрицы A .

Каждому собственному значению λ соответствует по крайней мере одно нетривиальное решение x . Такое решение называется *собственным вектором* или *характеристическим вектором*, соответствующим этому собственному значению. Если ранг матрицы $(A - \lambda I)$ меньше чем $(n - 1)$, то будет не менее двух линейно независимых векторов, удовлетворяющих уравнению (2.2). Мы детально обсудим этот вопрос в §§ 6—9. Очевидно, что если x — решение уравнения (2.2), то kx — тоже решение при любом значении k , так что, даже если ранг $(A - \lambda I)$ равен $(n - 1)$, собственный вектор, соответствующий λ , определен с точностью до произвольного множителя. Удобно выбирать этот множитель так, чтобы собственный вектор имел желаемые численные свойства; такие векторы называются *нормированными* векторами. Наиболее удобными способами нормировки являются те, при которых:

- (i) сумма квадратов модулей компонент равна единице;
- (ii) наибольшая по модулю компонента вектора равна единице;
- (iii) сумма модулей компонент равна единице.

Нормировки (i) и (iii) неудобны тем, что они не меняются при умножении вектора на комплексный множитель, по модулю равный единице.

Многие процессы, описанные в этой книге, приводят на первом этапе к ненормированным собственным векторам. Часто удобно умножать такие векторы на степени 10 (или 2) так, чтобы наибольшая компонента была по модулю между 0,1 и 1,0 (или 1/2 и 1). Такие векторы будем называть *стандартизованными* векторами.

Собственные значения и собственные векторы транспонированной матрицы

3. Собственные значения и собственные векторы транспонированной матрицы играют важную роль в общей теории. Собственные значения A^T это по определению те значения λ , при которых система уравнений

$$A^T y = \lambda y \quad (3.1)$$

имеет нетривиальное решение, т. е. те значения, для которых

$$\det(A^T - \lambda I) = 0. \quad (3.2)$$

Так как определитель матрицы равен определителю ее транспонированной, собственные значения A^T совпадают с собственными значениями A . Собственные векторы, вообще говоря, различны. Мы будем обычно обозначать через y_i собственный вектор A^T , соответствующий λ_i , так что

$$A^T y_i = \lambda_i y_i. \quad (3.3)$$

Уравнение (3.3) можно записать в виде

$$y_i^T A = \lambda_i y_i^T, \quad (3.4)$$

и потому y_i^T иногда называют *левым* собственным вектором A , соответствующим λ_i , тогда как вектор x_i такой, что

$$A x_i = \lambda_i x_i \quad (3.5)$$

называют *правым* собственным вектором, соответствующим λ_i . Важная роль собственных векторов транспонированной матрицы становится сразу ясной из следующего результата.

Если x_i — собственный вектор A , соответствующий λ_i , а y_j — собственный вектор A^T , соответствующий λ_j , то

$$x_j^T y_i = 0 \quad (\text{если } \lambda_i \neq \lambda_j), \quad (3.6)$$

Действительно, мы можем записать (3.5) в виде

$$x_i^T A^T = \lambda_i x_i^T, \quad (3.7)$$

а по определению

$$A^T y_j = \lambda_j y_j. \quad (3.8)$$

Умножая (3.7) справа на y_j и (3.8) слева на x_i^T и вычитая, получаем

$$0 = \lambda_i x_i^T y_j - \lambda_j x_i^T y_j, \quad (3.9)$$

что приводит к уравнению (3.6).

Заметим, что, так как x_i и y_j могут быть комплексными векторами, $x_i^T y_j$ не является скалярным произведением в обычном смысле. В самом деле мы имеем

$$x_i^T y_j = \overline{y_j^T x_i}, \quad (3.10)$$

а не

$$x_i^T y_j = \overline{(y_j^T x_i)}. \quad (3.14)$$

Если x — комплексный, может случиться, что

$$x^T x = 0, \quad (3.12)$$

в то время как настоящее скалярное произведение всегда положительно для всех ненулевых x .

Различные собственные значения

4. Общая теория проблемы собственных значений наиболее проста, когда все n собственных значений различны, и мы сначала рассмотрим этот случай. Обозначим собственные значения через $\lambda_1, \lambda_2, \dots, \lambda_n$. Уравнение (2.2) всегда имеет по крайней мере одно решение для каждого значения λ_i , и поэтому мы можем предполагать существование системы собственных векторов, которые мы обозначим через $x_1, x_2, \dots, x_n$. Покажем, что каждый из этих векторов единствен с точностью до произвольного множителя.

Сначала мы покажем, что эти векторы должны быть линейно независимыми. Действительно, пусть это не так, и пусть r — наименьшее число линейно зависимых векторов, которые мы занумеруем как $x_1, x_2, \dots, x_r$. По нашему предположению, между ними существует соотношение вида

$$\sum_1^r \alpha_i x_i = \vartheta, \quad (4.1)$$

в котором все α_i не равны нулю. Ясно, что $r \geq 2$, так как по определению все x_i — ненулевые векторы. Умножая (4.1) слева на A , получаем

$$\sum_1^r \alpha_i \lambda_i x_i = 0. \quad (4.2)$$

Умножая уравнение (4.1) на λ_r и вычитая (4.2), получим

$$\sum_1^{r-1} \alpha_i (\lambda_r - \lambda_i) x_i = 0, \quad (4.3)$$

причем ни один коэффициент в этом соотношении не равен нулю. Уравнение (4.3) означает, что $x_1, x_2, \dots, x_{r-1}$ линейно зависимы, что противоречит нашей гипотезе. Поэтому n собственных векторов линейно независимы, и все n -мерное пространство натянуто на них. Следовательно, они могут быть использованы в качестве базиса для представления произвольного вектора.

Отсюда единственность собственного вектора следует немедленно. Действительно, если существует второй собственный вектор z_1 , соответствующий λ_1 , мы можем написать

$$z_1 = \sum_1^n \alpha_i x_i, \quad (4.4)$$

где по крайней мере одно из α_i не равно нулю. Умножение (4.4) на A дает

$$\lambda_1 z_1 = \sum_1^n \alpha_i \lambda_i x_i. \quad (4.5)$$

Умножая (4.4) на λ_1 и вычитая из (4.5), получаем

$$0 = \sum_2^n \alpha_i (\lambda_i - \lambda_1) x_i, \quad (4.6)$$

и так как x_i линейно независимы,

$$\alpha_i (\lambda_i - \lambda_1) = 0 \quad (i = 2, 3, \dots, n),$$

что дает

$$\alpha_i = 0 \quad (i = 2, 3, \dots, n). \quad (4.7)$$

Поэтому ненулевым α_i должен быть α_1 , и значит, z_1 пропорционален x_1 .

Аналогично мы можем показать, что собственные векторы A^T единственны и линейно независимы. Но мы раньше доказали, что

$$x_i^T y_j = 0 \quad (\lambda_i \neq \lambda_j), \quad (4.8)$$

так что эти две системы векторов образуют *биортогональную систему*. Кроме того, имеем

$$x_i^T y_i \neq 0 \quad (i = 1, 2, \dots, n), \quad (4.9)$$

так как если x_i был бы ортогонален y_i , он был бы ортогонален к $y_1, y_2, \dots, y_n$ и, следовательно, ко всему n -мерному пространству, а это невозможно, так как x_i ненулевой вектор.

Уравнения (4.8) и (4.9) дают нам возможность получать явное выражение произвольного вектора v через x_i или y_i . Действительно, если мы напомним

$$v = \sum_{i=1}^n \alpha_i x_i, \quad (4.10)$$

то

$$y_j^T v = \sum_{i=1}^n \alpha_i y_j^T x_i = \alpha_j y_j^T x_j. \quad (4.11)$$

Следовательно,

$$\alpha_j = y_j^T v / y_j^T x_j, \quad (4.12)$$

причем деление возможно, так как знаменатель не равен нулю.

Преобразования подобия

5. Если мы выберем произвольные множители для x_i и y_j так, что

$$y_i^T x_i = 1 \quad (i = 1, 2, \dots, n), \quad (5.1)$$

то соотношения (4.8) и (5.1) будут означать, что матрица Y^T , имеющая y_i^T своей i -й строкой, обратна матрице X , у которой i -й столбец есть x_i . Все n систем уравнений

$$A x_i = \lambda_i x_i \quad (5.2)$$

могут быть записаны в матричной форме

$$A X = X \operatorname{diag}(\lambda_i). \quad (5.3)$$

Мы только что установили, что обратная матрица к матрице X существует и равна Y^T . Уравнение (5.3), следовательно, дает

$$X^{-1} A X = Y^T A X = \operatorname{diag}(\lambda_i). \quad (5.4)$$

Преобразование $H^{-1}AH$ матрицы A , где H неособенная, играет фундаментальную роль как с теоретической, так и с практической точки зрения, и оно известно как *преобразование подобия*, а матрицы A и $H^{-1}AH$ называются *подобными*. Очевидно, что и HAH^{-1} также является преобразованием подобия A . Мы показали, что если собственные значения матрицы A различны, то существует *преобразование подобия*, которое приводит матрицу A к диагональной форме, и что столбцы *матрицы преобразования* равны собственным векторам A . С другой стороны, если мы имеем

$$H^{-1}AH = \text{diag}(\mu_i), \quad (5.5)$$

то

$$AH = H \text{diag}(\mu_i). \quad (5.6)$$

Последнее уравнение означает, что числа μ_i — это собственные значения A , расположенные в некотором порядке, а i -й столбец H — это собственный вектор, соответствующий μ_i .

Собственные значения матрицы инвариантны при преобразовании подобия. Действительно, если

$$Ax = \lambda x, \quad (5.7)$$

то

$$H^{-1}Ax = \lambda H^{-1}x,$$

что дает

$$H^{-1}A(HH^{-1})x = \lambda H^{-1}x, \quad (H^{-1}AH)H^{-1}x = \lambda H^{-1}x. \quad (5.8)$$

Таким образом, собственные значения сохранились, а собственные векторы умножились на H^{-1} .

Многие численные методы нахождения собственных значений и собственных векторов матрицы по существу состоят в нахождении преобразования подобия, которое приводит матрицу A общего вида к матрице B специального вида, для которой проблема собственных значений может быть решена более просто.

Преобразование подобия транзитивно, именно, если

$$B = H_1^{-1}AH_1 \quad \text{и} \quad C = H_2^{-1}BH_2, \quad (5.9)$$

то

$$C = H_2^{-1}H_1^{-1}AH_1H_2 = (H_1H_2)^{-1}A(H_1H_2). \quad (5.10)$$

Приведение матрицы общего вида к одной из специальных форм будет, вообще говоря, осуществляться при помощи последовательности простых преобразований подобия.

Кратные собственные значения и канонические формы матриц общего вида

6. Структура системы собственных векторов матрицы, у которой одно или более собственных значений кратны, может быть далеко не так проста, как описано выше. Может все же оказаться, что существует преобразование подобия, которое приводит A к диагональной форме. Если это так, то для некоторой неособенной H мы имеем

$$H^{-1}AH = \text{diag}(\lambda_i). \quad (6.1)$$

Числа λ_i должны быть собственными значениями A , и каждое λ_i должно встречаться с соответствующей кратностью. Действительно,

$$H^{-1}(A - \lambda I)H = H^{-1}AH - \lambda H^{-1}IH = \text{diag}(\lambda_i - \lambda), \quad (6.2)$$

и, беря определитель от обеих частей, получаем

$$\det(H^{-1}) \det(A - \lambda I) \det(H) = \prod (\lambda_i - \lambda), \quad (6.3)$$

что дает

$$\det(A - \lambda I) = \prod (\lambda_i - \lambda). \quad (6.4)$$

Поэтому λ_i суть корни характеристического уравнения A . Записав (6.1) в виде

$$AH = H \text{diag}(\lambda_i), \quad (6.5)$$

мы видим, что столбцы H являются собственными векторами A . Так как H неособенная, ее столбцы линейно независимы. Если, например, λ_1 — двукратный корень, то, обозначая первые два столбца через x_1 и x_2 , получим

$$Ax_1 = \lambda_1 x_1, \quad Ax_2 = \lambda_1 x_2, \quad (6.6)$$

где x_1 и x_2 линейно независимы. Из уравнений (6.6) следует, что любой вектор из подпространства, натянутого на x_1 и x_2 , также является собственным вектором. Действительно,

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 \lambda_1 x_1 + \alpha_2 \lambda_1 x_2 = \lambda_1(\alpha_1 x_1 + \alpha_2 x_2). \quad (6.7)$$

Совпадение корней у матрицы, которую можно привести к диагональной форме преобразованием подобия, приводит к неопределенности собственных векторов, соответствующих кратному корню. Мы все же можем выбрать в этом случае систему собственных векторов, на которые натянуто все n -мерное пространство и которые можно использовать в качестве базиса для представления произвольного вектора.

Так, диагональная матрица

$$\Lambda = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 2 & & \\ & & & 2 & \\ & & & & 3 \end{bmatrix} \quad (6.8)$$

имеет пять линейно независимых собственных векторов e_1, e_2, e_3, e_4 и e_5 . Любая линейная комбинация e_1 и e_2 — это собственный вектор, соответствующий $\lambda = 1$, и любая линейная комбинация e_3 и e_4 — собственный вектор, соответствующий $\lambda = 2$. Для каждой матрицы, подобной Λ , существует неособенная H такая, что

$$H^{-1}AH = \Lambda \quad (6.9)$$

и собственные векторы A суть He_1, He_2, He_3, He_4 и He_5 . Любая линейная комбинация первых двух будет собственным вектором, соответствующим $\lambda = 1$, и любая линейная комбинация третьего и четвертого — собственным вектором, соответствующим $\lambda = 2$.

Неполная система собственных векторов

7. Свойства матриц, подобных диагональной, но имеющих кратные корни, очень сходны со свойствами матриц с различными собственными значениями. Однако не все матрицы с кратными собственными значениями таковы.

Матрица

$$C_2(a) = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix} \quad (7.1)$$

имеет двукратное собственное значение $\lambda = a$. Если x — собственный вектор $C_2(a)$ с компонентами x_1 и x_2 , то

$$\left. \begin{aligned} 0x_1 + x_2 &= 0, \\ 0x_1 + 0x_2 &= 0. \end{aligned} \right\} \quad (7.2)$$

Эти уравнения означают, что $x_2 = 0$, так что существует лишь один собственный вектор, соответствующий $\lambda = a$, именно вектор e_1 . Мы можем рассматривать $C_2(a)$ как предел при $b \rightarrow a$ матрицы B_2 , равной

$$B_2 = \begin{bmatrix} a & 1 \\ 0 & b \end{bmatrix}. \quad (7.3)$$

При $b \neq a$ эта матрица имеет два собственных значения a и b с соответствующими векторами

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} 1 \\ b - a \end{bmatrix}.$$

При $b \rightarrow a$ оба собственных значения и оба собственных вектора совпадают. Если мы определим матрицы $C_r(a)$ соотношениями

$$C_1(a) = [a], \quad C_r(a) = \begin{bmatrix} a & 1 & & & \\ & a & 1 & & \\ & & a & 1 & \\ \dots & \dots & \dots & \dots & \dots \\ & & & a & 1 \\ & & & & a \end{bmatrix}, \quad (7.4)$$

то $C_r(a)$ будет иметь единственное собственное значение a кратности r , и только один собственный вектор $x = e_1$. Мы можем доказать это так же, как и для $C_2(a)$. С другой стороны, это следует из того, что ранг $[C_r(a) - aI]$ равен $r - 1$. (Определитель матрицы порядка $r - 1$ в правом верхнем углу равен единице). Транспонированная матрица $[C_r(a)]^T$ также будет иметь лишь один собственный вектор $y = e_r$. Очевидно,

$$x^T y = e_1^T e_r = 0, \quad (7.5)$$

что отличается от результата для матриц с различными собственными значениями.

Матрицу $C_r(a)$ ($r > 1$) нельзя привести к диагональной форме преобразованием подобия. Действительно, если существует H такая, что

$$H^{-1} C_r H = \text{diag}(\lambda_i), \quad (7.6)$$

т. е.

$$C_r H = H \text{diag}(\lambda_i), \quad (7.7)$$

то, как показано в § 6, λ_i должны совпадать с собственными значениями $C_r(a)$ и, следовательно, должны быть все равны a . Из уравнения (7.7) тогда следует, что все столбцы матрицы H — это собственные векторы $C_r(a)$. Эти столбцы линейно независимы, так как H неособенная. Предположение о существовании такой H , следовательно, ложно.

Матрица $C_r(a)$ называется *простой жордановой подматрицей* порядка r или иногда *простой классической подматрицей* порядка r . Матрицы этого вида играют важную роль в теории, как будет видно в следующем параграфе.

Жорданова (классическая) каноническая форма

8. Мы видим, что матрица с кратными собственными значениями не обязательно подобна диагональной матрице. Естествен вопрос, к какой наиболее компактной форме может быть приведена матрица A с помощью преобразования подобия. Ответ содержится в следующей теореме.

Пусть A — матрица порядка n с r различными собственными значениями $\lambda_1, \lambda_2, \dots, \lambda_r$ кратностей $m_1, m_2, \dots, m_r$ таких, что

$$m_1 + m_2 + \dots + m_r = n. \quad (8.1)$$

Тогда существует преобразование подобия с матрицей H такое, что матрица $H^{-1}AH$ состоит из простых жордановых подматриц, размещенных вдоль диагонали, а все ее остальные элементы равны нулю. Сумма порядков подматриц, связанных с λ_i , равна m_i . С точностью до порядка размещения подматриц вдоль диагонали преобразующая матрица единственна. Будем говорить, что $H^{-1}AH$ есть прямая сумма простых жордановых подматриц.

Так, матрица шестого порядка с пятикратным собственным значением λ_1 и однократным λ_2 , подобная матрице

$$C = \begin{bmatrix} C_2(\lambda_1) & & & \\ & C_2(\lambda_1) & & \\ & & C_1(\lambda_1) & \\ & & & C_1(\lambda_2) \end{bmatrix} \quad (8.2)$$

не может быть также подобна матрице вида

$$\begin{bmatrix} C_3(\lambda_1) & & \\ & C_2(\lambda_1) & \\ & & C_1(\lambda_2) \end{bmatrix}, \quad (8.3)$$

хотя эта матрица также имеет пять корней, равных λ_1 , и один, равный λ_2 .

Матрица, состоящая из простых жордановых подматриц, называется *жордановой канонической формой* или *классической канонической формой* матрицы A .

В частном случае, когда все жордановы подматрицы первого порядка, эта форма становится диагональной. Мы уже видели, что если все собственные значения матрицы различны, то ее жорданова каноническая форма есть диагональная матрица $\text{diag}(\lambda_i)$.

Общее число линейно независимых собственных векторов равно числу подматриц в жордановой канонической матрице. Так, матрица A , которая может быть приведена к C , определенной в (8.2), имеет четыре линейно независимых собственных вектора. Собственные векторы C — это e_1, e_3, e_5 и e_6 и, следовательно, собственные векторы A — He_1, He_3, He_5

и He_6 . Собственному значению λ_1 , кратность которого равна пяти, соответствуют только три собственных вектора.

Не будем доказывать существование и единственность жордановой канонической формы, так как метод доказательства имеет мало общего с методами, описанными в последующих главах. Отметим только, что в простых жордановых подматрицах порядка r мы взяли элементы, равные единице, лежащими над диагональю. Существует соответствующая форма, когда эти элементы взяты лежащими под диагональю. Будем называть форму (7.4) *верхней* жордановой формой, а другую форму — *нижней* жордановой формой. Выбор значения единица для этих элементов не играет особой роли. Ясно, что подходящим преобразованием подобия с диагональной матрицей эти элементы можно сделать произвольными величинами, не равными нулю.

Элементарные делители

9. Пусть C — жорданова каноническая форма A . Рассмотрим матрицу $(C - \lambda I)$. Для матрицы C , определенной в (8.2), например, имеем

$$(C - \lambda I) = \begin{bmatrix} C_2(\lambda_1 - \lambda) & & & \\ & C_2(\lambda_1 - \lambda) & & \\ & & C_1(\lambda_1 - \lambda) & \\ & & & C_1(\lambda_2 - \lambda) \end{bmatrix}. \quad (9.1)$$

Ясно, что $(C - \lambda I)$ есть прямая сумма матриц вида $C_r(\lambda_i - \lambda)$. Определители подматриц канонической формы $(C - \lambda I)$ называются *элементарными делителями* матрицы A . Так, элементарными делителями любой матрицы, подобной C из (8.2), будут

$$(\lambda_1 - \lambda)^{2^r}, (\lambda_1 - \lambda)^2, (\lambda_1 - \lambda) \quad \text{и} \quad (\lambda_2 - \lambda). \quad (9.2)$$

Очевидно, что произведение элементарных делителей матрицы равно ее характеристическому полиному. Если жорданова каноническая форма диагональна, то элементарными делителями будут

$$(\lambda_1 - \lambda), (\lambda_2 - \lambda), \dots, (\lambda_n^1 - \lambda), \quad (9.3)$$

так что все они линейны. У матрицы с различными собственными значениями ее элементарные делители всегда линейны, но мы видели, что если у матрицы есть одно или более кратное собственное значение, то она может иметь *линейные элементарные делители*, а может и не иметь.

Если у матрицы есть один или более *нелинейный* элементарный делитель, то по крайней мере одна из подматриц в ее жордановой канонической форме имеет порядок, больший единицы, и, следовательно, у нее будет меньше чем n линейно независимых собственных векторов. Такие матрицы называются *недостаточными* *).

Сопровождающая матрица характеристического полинома

10. Обозначим характеристическое уравнение через

$$(-1)^n [\lambda^n - p_{n-1}\lambda^{n-1} - p_{n-2}\lambda^{n-2} - \dots - p_0] = 0. \quad (10.1)$$

*) Термин автора defective не устремителен в русской литературе. Матрицу с линейными элементарными делителями называют матрицей простой структуры. (Прим. перев.)

Легко проверить, что характеристическое уравнение матрицы

$$C = \begin{bmatrix} p_{n-1} & p_{n-2} & \dots & p_1 & p_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad (10.2)$$

совпадает с характеристическим уравнением A . Матрица C называется *сопровождающей матрицей* характеристического полинома матрицы A . Так как у C и A одинаковые характеристические полиномы, естественно задать вопрос: когда они подобны? Мы дадим ответ на этот вопрос в следующем параграфе. Заметим, что мы могли бы задавать сопровождающую матрицу в трех других формах. Например, в форме

$$\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & p_0 \\ 1 & 0 & 0 & \dots & 0 & p_1 \\ 0 & 1 & 0 & \dots & 0 & p_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & p_{n-1} \end{bmatrix}. \quad (10.3)$$

Элементы p_i можно было бы расположить также в первом столбце или в последней строке, а единичные элементы — над диагональю. Мы будем использовать ту из этих форм сопровождающей матрицы, которая будет наиболее удобна.

Полные матрицы

11. Предположим, что собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$ матрицы A различны. Тогда, как известно, A подобна $\text{diag}(\lambda_i)$. Покажем, что в этом случае $\text{diag}(\lambda_i)$ подобна сопровождающей матрице. Мы докажем это, указав матрицу, которая преобразует C в $\text{diag}(\lambda_i)$. Проиллюстрируем общий случай, рассмотрев случай $n = 4$.

Рассмотрим матрицу

$$H = \begin{bmatrix} \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \quad (11.1)$$

Если все λ_i различны, эта матрица неособенная, и мы имеем

$$H \text{diag}(\lambda_i) = \begin{bmatrix} \lambda_1^4 & \lambda_2^4 & \lambda_3^4 & \lambda_4^4 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}. \quad (11.2)$$

Далее, соотношение

$$\begin{bmatrix} p_3 & p_2 & p_1 & p_0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} H = \begin{bmatrix} \lambda_1^4 & \lambda_2^4 & \lambda_3^4 & \lambda_4^4 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix} \quad (11.3)$$

следует из того, что

$$\lambda_i^4 = p_3 \lambda_i^3 + p_2 \lambda_i^2 + p_1 \lambda_i + p_0. \quad (11.4)$$

Следовательно,

$$CH = H \operatorname{diag}(\lambda_i), \quad (11.5)$$

так что C подобна $\operatorname{diag}(\lambda_i)$ и, следовательно, любой матрице с собственными значениями λ_i .

Возвращаясь к случаю нескольких кратных собственных значений, мы сразу видим, что если матрица A имеет два или более линейно независимых собственных вектора, соответствующих λ_i , то она не может быть подобна C .

Действительно, ранг $(C - \lambda I)$ равен $n - 1$ при любом значении λ , так как определитель матрицы порядка $n - 1$ в нижнем левом углу, очевидно, равен единице. Собственный вектор x_i , соответствующий собственному значению λ_i , равен

$$x_i^T = [\lambda_i^{n-1}, \lambda_i^{n-2}, \dots, \lambda_i, 1], \quad (11.6)$$

что можно увидеть, решая систему

$$(C - \lambda_i I) x_i = 0. \quad (11.7)$$

Для того чтобы матрица была подобна сопровождающей матрице своего характеристического полинома, необходимо, чтобы в ее каноническую жорданову форму входила только одна жорданова подматрица, соответствующая каждому различному собственному значению. Сейчас мы докажем, что это условие также и достаточно и снова приведем доказательство лишь для некоторого простого случая.

12. Рассмотрим матрицу четвертого порядка, элементарные делители которой равны $(\lambda_1 - \lambda)^2, (\lambda_3 - \lambda), (\lambda_4 - \lambda)$. Такая матрица подобна матрице A_4 , равной

$$A_4 = \begin{bmatrix} \lambda_1^3 & 1 & 0 & 0 \\ 0 & \lambda_1^3 & 0 & 0 \\ & & \lambda_3 & 0 \\ & & & \lambda_4 \end{bmatrix}. \quad (12.1)$$

Матрица

$$H = \begin{bmatrix} \lambda_1^3 & 3\lambda_1^2 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & 1 & \lambda_3 & \lambda_4 \\ 1 & 0 & 1 & 1 \end{bmatrix} \quad (12.2)$$

очевидно, неособенная, если $\lambda_1, \lambda_3, \lambda_4$ различны, и мы имеем

$$HA_4 = \begin{bmatrix} \lambda_1^4 & 4\lambda_1^3 & \lambda_3^4 & \lambda_4^4 \\ \lambda_1^3 & 3\lambda_1^2 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & 1 & \lambda_3 & \lambda_4 \end{bmatrix}. \quad (12.3)$$

Так как λ_1 — двукратный корень характеристического уравнения $f(\lambda) = 0$, то $f'(\lambda_1) = 0$, что дает

$$4\lambda_1^3 = 3p_3\lambda_1^2 + 2p_2\lambda_1 + p_1, \quad (12.4)$$

и, следовательно,

$$CH = HA_i. \quad (12.5)$$

Аналогичным образом мы можем поступить с жордановой подматрицей порядка r , связанной с матрицей порядка n . Если λ_1 — соответствующее собственное значение, то r столбцов матрицы H имеют вид

$$\begin{bmatrix} \lambda_1^{n-1} & \binom{n-1}{1} \lambda_1^{n-2} & \dots & \binom{n-1}{r-1} \lambda_1^{n-r} \\ \lambda_1^{n-2} & \binom{n-2}{1} \lambda_1^{n-3} & \dots & \binom{n-2}{r-1} \lambda_1^{n-r-1} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_1 & 1 & \dots & 0 \\ 1 & 0 & \dots & 0 \end{bmatrix}. \quad (12.6)$$

Таким образом, предполагая, что каждому различному собственному значению соответствует лишь одна жорданова подматрица, и, следовательно, лишь один собственный вектор, мы получим, что матрица подобна сопровождающей матрице своего характеристического полинома. Такие матрицы называются *полными*.

Каноническая (рациональная) форма Фробениуса

13. Матрица называется *неполной*, если с λ_i при некотором i связана более чем одна жорданова подматрица (и поэтому более чем один собственный вектор). Мы видели, что неполная матрица не может быть подобна сопровождающей матрице своего характеристического полинома. Найдем для неполных матриц каноническую матрицу, аналогичную сопровождающей матрице.

Определим *матрицу Фробениуса* B_r порядка r

$$B_r = \begin{bmatrix} b_{r-1} & b_{r-2} & \dots & b_1 & b_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \quad (13.1)$$

Фундаментальная теорема для матриц общего вида может быть сформулирована так.

Каждая матрица преобразованием подобия может быть приведена к прямой сумме s матриц Фробениуса, которые можно обозначить B_{r_1} , B_{r_2} , ..., B_{r_s} . Характеристический полином каждой B_{r_i} делит характеристические полиномы всех предыдущих B_{r_j} . В случае полной матрицы $s = 1$ и $r_1 = n$. Прямая сумма матриц Фробениуса называется канонической формой *Фробениуса* (или *рациональной*). Наименование рациональной канонической формы возникло из того факта, что она может быть получена преобразованиями, рациональными в поле элементов A .

Мы не будем давать прямое доказательство существования канонической формы Фробениуса, но укажем ее связь с жордановой канонической формой. Это не проиллюстрирует рациональную природу преобразования, так как операции, дающие жорданову форму, вообще говоря, иррациональны в поле A .

Связь между жордановой и фробениусовой каноническими формами

14. Связь между двумя каноническими формами, вероятно, лучше всего проиллюстрировать на примере. Пусть матрица A десятого порядка имеет следующие элементарные делители:

$$\left. \begin{aligned} &(\lambda_1 - \lambda)^3, \quad (\lambda_2 - \lambda)^2, \quad (\lambda_3 - \lambda), \\ &(\lambda_1 - \lambda), \quad (\lambda_2 - \lambda)^2, \\ &(\lambda_1 - \lambda). \end{aligned} \right\} \quad (14.1)$$

Ее жорданову каноническую форму можно записать в виде

$$\left[\begin{array}{ccc|ccc|ccc} C_3(\lambda_1) & & & & & & & & \\ & C_2(\lambda_2) & & & & & & & \\ & & C_1(\lambda_3) & & & & & & \\ \hline & & & C_1(\lambda_1) & & & & & \\ & & & & C_2(\lambda_2) & & & & \\ & & & & & C_2(\lambda_2) & & & \\ \hline & & & & & & C_1(\lambda_1) & & \end{array} \right] = \left[\begin{array}{ccc|ccc} G_1 & & & & & \\ & G_2 & & & & \\ & & G_3 & & & \\ \hline & & & G_1 & & \\ & & & & G_2 & \\ & & & & & G_3 \end{array} \right], \quad (14.2)$$

где мы сгруппировали жордановы подматрицы высшего порядка, соответствующие каждому собственному значению, в G_1 , следующего высшего порядка — в G_2 и т. д.

Каждая матрица G_i , следовательно, имеет только одну жорданову подматрицу, соответствующую каждому λ_i , и поэтому может быть приведена преобразованием подобия H_i к единственной матрице Фробениуса соответствующего порядка. В нашем примере

$$\left. \begin{aligned} H_1^{-1} G_1 H_1 &= B_6 \quad (\text{матрица Фробениуса 6-го порядка}), \\ H_2^{-1} G_2 H_2 &= B_3 \quad (\text{матрица Фробениуса 3-го порядка}), \\ H_3^{-1} G_3 H_3 &= B_1 \quad (\text{матрица Фробениуса 1-го порядка}). \end{aligned} \right\} \quad (14.3)$$

Следовательно,

$$\left[\begin{array}{ccc} H_1^{-1} & & \\ & H_2^{-1} & \\ & & H_3^{-1} \end{array} \right] \left[\begin{array}{ccc} G_1 & & \\ & G_2 & \\ & & G_3 \end{array} \right] \left[\begin{array}{ccc} H_1 & & \\ & H_2 & \\ & & H_3 \end{array} \right] = \left[\begin{array}{ccc} B_6 & & \\ & B_3 & \\ & & B_1 \end{array} \right]. \quad (14.4)$$

Характеристический полином B_6 равен $(\lambda_1 - \lambda)^3 (\lambda_2 - \lambda)^2 (\lambda_3 - \lambda)$, B_3 равен $(\lambda_1 - \lambda) (\lambda_2 - \lambda)^2$ и B_1 равен $(\lambda_1 - \lambda)$. Каждый из этих полиномов является делителем предшествующего, что следует из определения G_i и что остается справедливым в общем случае. Очевидно, что элементарные делители матрицы являются множителями в характеристических полиномах подматриц в канонической форме Фробениуса. Заметим, что если матрица A вещественна, то каждое комплексное собственное значение встречается как сопряженная пара. Следовательно, каждая B_i будет вещественная. Этого можно было ожидать, так как мы утверждали, что каноническая форма Фробениуса может быть получена преобразованиями, рациональными в поле A .

Преобразования эквивалентности

15. Мы видели, что собственные значения матрицы инвариантны по отношению к преобразованиям подобия. Рассмотрим сейчас кратко более широкий класс преобразований. Назовем матрицу B *эквивалентной* матрице A , если существуют неособенные (обязательно квадратные) матрицы

P и Q такие, что

$$PAQ = B. \quad (15.1)$$

Заметим, что A и B не обязательно должны быть квадратными, но они должны быть одинаковых размеров. Из теоремы об определителе произведения двух прямоугольных матриц (иногда известной как теорема Бине — Коши; см., например, Гантмахер, 1967) следует, что эквивалентные матрицы имеют одинаковый ранг.

Отношение эквивалентности, которое мы сейчас рассмотрели, существенно при решении систем линейных уравнений. Действительно, если x удовлетворяет уравнению

$$Ax = b, \quad (15.2)$$

то ясно, что x удовлетворяет и уравнению

$$PAx = Pb \quad (15.3)$$

при любой матрице P . Если, далее, P — неособенная, то любое решение (15.3) есть решение (15.2), что можно увидеть, умножая (15.3) на P^{-1} .

Теперь рассмотрим замену переменных

$$x = Qy, \quad (15.4)$$

где Q — неособенная. Тогда

$$PAQy = Pb, \quad (15.5)$$

и из любого решения y (15.5) можем получить решение x (15.2), и обратно. Уравнения (15.2) и (15.5) могут быть, следовательно, рассмотрены как эквивалентные.

λ -матрицы

16. Расширим концепцию эквивалентности на матрицы, элементами которых являются полиномы от λ . Такие матрицы называются λ -матрицами. Если $A(\lambda)$ — $(m \times n)$ -матрица *), элементы которой суть полиномы степени не выше k , то можно записать

$$A(\lambda) = A_k \lambda^k + A_{k-1} \lambda^{k-1} + \dots + A_0, \quad (16.1)$$

где $A_k, A_{k-1}, \dots, A_0$ — $(m \times n)$ -матрицы, элементы которых не зависят от λ . Ранг λ -матрицы — это максимальный порядок миноров, которые не являются нулевыми полиномами от λ .

Две λ -матрицы $A(\lambda)$ и $B(\lambda)$ называются эквивалентными, если существуют квадратные λ -матрицы $P(\lambda)$ и $Q(\lambda)$ с не зависящими от λ ненулевыми определителями такие, что

$$P(\lambda) A(\lambda) Q(\lambda) = B(\lambda). \quad (16.2)$$

Заметим, что так как определитель $P(\lambda)$ отличен от нуля и не зависит от λ , $[P(\lambda)]^{-1}$ есть также λ -матрица с теми же свойствами.

Элементарные операции

17. Нас будут интересовать простые неособенные λ -матрицы $P(\lambda)$ следующих типов:

- (i) $p_{ii}(\lambda) = 1$ ($i \neq p, q$), $p_{pq}(\lambda) = p_{qp}(\lambda) = 1$, остальные $p_{ij}(\lambda) = 0$;
- (ii) $p_{ii}(\lambda) = 1$ ($i = 1, \dots, n$), $p_{pq}(\lambda) = f(\lambda)$, остальные $p_{ij}(\lambda) = 0$;
- (iii) $p_{ii}(\lambda) = 1$ ($i \neq p$), $p_{pp}(\lambda) = k \neq 0$, остальные $p_{ij}(\lambda) = 0$.

*) То есть матрица, состоящая из m строк и n столбцов. (Прим. перев.)

Очевидно, что определитель матрицы типа (i) равен -1 , типа (ii) равен $+1$ и типа (iii) равен k .

При умножении матрицы $A(\lambda)$ слева на матрицу типа (i), (ii) или (iii) соответственно:

- (i) переставляются строки p и q ;
- (ii) к строке p прибавляется строка q , умноженная на $f(\lambda)$;
- (iii) строка p умножается на отличную от нуля постоянную k .

Умножение справа производит соответствующие операции со столбцами. Назовем такие преобразования $A(\lambda)$ *элементарными операциями*.

Каноническая форма Смита

18. Фундаментальная теорема об эквивалентных λ -матрицах состоит в следующем.

Каждая λ -матрица $A(\lambda)$ ранга r может быть приведена элементарными операциями (рациональными в поле элементов $A(\lambda)$) к эквивалентной диагональной форме вида

$$D = \begin{bmatrix} E_1(\lambda) & & & & \\ & E_2(\lambda) & & & \\ & & \ddots & & \\ & & & E_r(\lambda) & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix}, \quad (18.1)$$

где каждый $E_i(\lambda)$ — это *нормализованный* полином такой, что $E_i(\lambda)$ делит $E_{i+1}(\lambda)$. (Старший коэффициент нормализованного полинома равен единице.) Все остальные элементы D равны нулю.

Доказательство таково. Будем получать диагональную форму с помощью последовательных элементарных операций, обозначая (i, j) -й элемент *текущих* преобразованных матриц на каждом шаге через $a_{ij}(\lambda)$.

Пусть $a_{ij}(\lambda)$ — один из элементов $A(\lambda)$ наименьшей степени. Перестановкой первой и i -й строк и первого и j -го столбцов этот элемент может быть помещен в позицию $(1,1)$. Пусть

$$\left. \begin{aligned} a_{i1}(\lambda) &= a_{11}(\lambda)q_{i1}(\lambda) + r_{i1}(\lambda) & (i = 2, 3, \dots, m), \\ a_{1j}(\lambda) &= a_{11}(\lambda)q_{1j}(\lambda) + r_{1j}(\lambda) & (j = 2, 3, \dots, n), \end{aligned} \right\} \quad (18.2)$$

где $r_{i1}(\lambda)$ и $r_{1j}(\lambda)$ — остатки, так что их степень меньше, чем степень $a_{11}(\lambda)$. Если не все $r_{i1}(\lambda)$ и $r_{1j}(\lambda)$ равны нулю, то предположим, что $r_{k1}(\lambda)$ не нуль. Вычтем из k -й строки первую, умноженную на $q_{k1}(\lambda)$, и переставим первую и k -ю строку. В позиции $(1,1)$ тогда будет остаток $r_{k1}(\lambda)$, отличный от нуля и степени меньшей по сравнению с предыдущим элементом на этом месте.

Продолжая таким образом, получим, что элемент $a_{11}(\lambda)$ остается отличным от нуля, и его степень постоянно уменьшается. В конечном счете $a_{11}(\lambda)$ должен делить все $a_{i1}(\lambda)$ и $a_{1j}(\lambda)$, либо когда $a_{11}(\lambda)$ уже не зависит от λ , либо на более ранней стадии. Когда это произошло, матрицу можно преобразовать к виду

$$\begin{bmatrix} a_{11}(\lambda) & 0 \\ 0 & A_2(\lambda) \end{bmatrix}, \quad (18.3)$$

вычитая из всех строк, начиная со второй, первую, умноженную на соответствующие множители, и затем вычитая из всех столбцов, начиная со второго, первый столбец, умноженный на соответствующие множители.

На этой стадии $a_{11}(\lambda)$ может делить все элементы $A_2(\lambda)$. Если это не так, положим для некоторого элемента $a_{ij}(\lambda)$

$$a_{ij}(\lambda) = a_{11}(\lambda) q_{ij}(\lambda) + r_{ij}(\lambda) \quad (r_{ij}(\lambda) \neq 0). \quad (18.4)$$

Теперь, добавляя i -ю строку к первой, мы можем применить предыдущую процедуру и получить форму (18.3) снова с ненулевым $a_{11}(\lambda)$ более низкой степени. Продолжая подобным образом, мы достигнем вида (18.3), где $a_{11}(\lambda)$ делит все элементы $A_2(\lambda)$. Это случится, либо когда $a_{11}(\lambda)$ станет уже независимым от λ , либо на более ранней стадии.

Если $A_2(\lambda)$ не тождественный нуль, мы можем применить к этой матрице процесс, подобный примененному к $A(\lambda)$. При этом мы можем полагать, что операции производятся лишь над последними $(m-1)$ строками и последними $(n-1)$ столбцами матрицы (18.3); первая строка и первый столбец уже не меняются. На всех стадиях этого процесса все элементы $A_2(\lambda)$ остаются кратными $a_{11}(\lambda)$. Таким образом, мы получили форму

$$\left[\begin{array}{ccc|c} a_{11}(\lambda) & 0 & & \\ 0 & a_{22}(\lambda) & & 0 \\ \hline & 0 & & A_3(\lambda) \end{array} \right], \quad (18.5)$$

в которой $a_{11}(\lambda)$ делит $a_{22}(\lambda)$, и $a_{22}(\lambda)$ делит все элементы $A_3(\lambda)$. Продолжая подобным образом, получаем форму

$$\left[\begin{array}{cccc|c} a_{11}(\lambda) & & & & \\ & a_{22}(\lambda) & & & \\ & & \ddots & & \\ & & & a_{ss}(\lambda) & 0 \\ \hline & 0 & & & 0 \end{array} \right], \quad (18.6)$$

где каждый элемент $a_{ii}(\lambda)$ делит все следующие, причем процесс продолжается до тех пор, пока либо s равно m или n , либо A_{s+1} — нулевая матрица. Однако во всяком случае $s = r$, так как по теореме Бине — Коши ранг λ -матрицы не меняется при умножении на неособенную матрицу. Так как $a_{ii}(\lambda)$ — отличные от нуля полиномы, то, умножая строки на соответствующие постоянные, можно сделать коэффициенты при старших степенях в каждом полиноме равными единице. Тогда $a_{ii}(\lambda)$ станут нормализованными полиномами $E_i(\lambda)$, и каждый $E_i(\lambda)$ делит $E_{i+1}(\lambda)$. Это завершает доказательство, причем из доказательства ясно, что использовались лишь рациональные операции. Так как D получена элементарными операциями, то

$$P(\lambda) A(\lambda) Q(\lambda) = D, \quad (18.7)$$

причем $P(\lambda)$ и $Q(\lambda)$ — неособенные матрицы с определителем, не зависящим от λ . Форма D (18.1) известна как каноническая форма Смита для эквивалентных λ -матриц.

Наибольший общий делитель k -строчных миноров λ -матрицы

19. Рассмотрим k -строчные миноры канонической формы Смита. Если $k > r$, то все k -строчные миноры равны нулю. С другой стороны, единственные отличные от нуля k -строчные миноры — это произведения k сомножителей, взятые из $E_i(\lambda)$. Так как каждый E_i есть множитель E_{i+1} , то наибольший общий делитель (н. о. д.) $G_k(\lambda)$ k -строчных миноров будет равен $E_1(\lambda) E_2(\lambda) \dots E_r(\lambda)$. Следовательно, если $G_0(\lambda) = 1$,

$$E_i(\lambda) = \frac{G_i(\lambda)}{G_{i-1}(\lambda)} \quad (i = 1, 2, \dots, r). \quad (19.1)$$

Мы покажем, что н. о. д. для k -строчных миноров матрицы (мы выберем коэффициент при высшей степени λ равным единице) инвариантен при преобразованиях эквивалентности. Действительно, если

$$P(\lambda) A(\lambda) Q(\lambda) = B(\lambda), \quad (19.2)$$

то любой k -строчный минор $B(\lambda)$ может быть выражен в виде суммы членов вида $P_\alpha A_\beta Q_\gamma$, где $P_\alpha, A_\beta, Q_\gamma$ означают k -строчные миноры $P(\lambda), A(\lambda), Q(\lambda)$ соответственно. Следовательно, н. о. д. для k -строчных миноров $A(\lambda)$ делит любой k -строчный минор $B(\lambda)$. С другой стороны, так как

$$A(\lambda) = [P(\lambda)]^{-1} B(\lambda) [Q(\lambda)]^{-1},$$

и $[P(\lambda)]^{-1}$ и $[Q(\lambda)]^{-1}$ также λ -матрицы, н. о. д. k -строчных миноров $B(\lambda)$ делит любой k -строчный минор $A(\lambda)$. Поэтому два н. о. д. должны совпадать.

Все эквивалентные λ -матрицы поэтому имеют одинаковую каноническую форму Смита. Полиномы $E_i(\lambda)$ называются *инвариантными множителями* или *инвариантными полиномами* $A(\lambda)$.

Инвариантные множители $(A - \lambda I)$

20. Если A — квадратная матрица, то $(A - \lambda I)$ — λ -матрица ранга n и, следовательно, имеет n инвариантных множителей $E_1(\lambda), E_2(\lambda), \dots, E_n(\lambda)$, каждый из которых делит последующие. Свяжем инвариантные множители $(A - \lambda I)$ с определителями матриц Фробениуса в канонической форме Фробениуса. Пусть B — каноническая форма Фробениуса матрицы A , так что существует H такая, что

$$H^{-1} A H = B, \quad (20.1)$$

и следовательно,

$$H^{-1} (A - \lambda I) H = B - \lambda I. \quad (20.2)$$

Так как матрицы H^{-1} и H неособенные и не зависят от λ , $B - \lambda I$ эквивалентна $A - \lambda I$.

Рассмотрим инвариантные множители $(B - \lambda I)$. Их общий вид достаточно наглядно виден на простом примере. Пусть $(B - \lambda I)$ имеет вид

$$\left[\begin{array}{cccc|ccc|cc} p_3 - \lambda & p_2 & p_1 & p_0 & & & & & & \\ 1 & -\lambda & 0 & 0 & & & & & & \\ 0 & 1 & -\lambda & 0 & & & & & & \\ 0 & 0 & 1 & -\lambda & & & & & & \\ \hline & & & & q_2 - \lambda & q_1 & q_0 & & & \\ & & & & 1 & -\lambda & 0 & & & \\ & & & & 0 & 1 & -\lambda & & & \\ \hline & & & & & & & r_1 - \lambda & r_0 & \\ & & & & & & & 1 & -\lambda & \end{array} \right] = \begin{bmatrix} B_1 - \lambda I & & \\ & B_2 - \lambda I & \\ & & B_3 - \lambda I \end{bmatrix}. \quad (20.3)$$

Из § 14 мы знаем, что в последовательности $\det(B_1 - \lambda I)$, $\det(B_2 - \lambda I)$, $\det(B_3 - \lambda I)$ каждый полином делится на последующие.

Перебирая ненулевые миноры, мы видим, что $G_i(\lambda)$, н. о. д. для i -строчных миноров, таковы:

$$\left. \begin{aligned} G_0(\lambda) &= 1 \text{ (по определению)}, G_1(\lambda) = G_2(\lambda) = \dots = G_6(\lambda) = 1, \\ G_7(\lambda) &= \det(B_3 - \lambda I), G_8(\lambda) = \det(B_3 - \lambda I) \det(B_2 - \lambda I), \\ G_9(\lambda) &= \det(B_3 - \lambda I) \det(B_2 - \lambda I) \det(B_1 - \lambda I). \end{aligned} \right\} \quad (20.4)$$

Следовательно, инвариантные множители равны

$$\left. \begin{aligned} E_1(\lambda) &= E_2(\lambda) = \dots = E_6(\lambda) = 1, E_7(\lambda) = \det(B_3 - \lambda I), \\ E_8(\lambda) &= \det(B_2 - \lambda I), E_9(\lambda) = \det(B_1 - \lambda I). \end{aligned} \right\} \quad (20.5)$$

Очевидно, что наш результат легко обобщается, так что инвариантные множители $(A - \lambda I)$, не равные единице, равны характеристическим полиномам подматриц в канонической форме Фробениуса. Основной результат относительно подобия и λ -эквивалентности следующий.

Для того чтобы A была подобна B , необходимо и достаточно, чтобы $(A - \lambda I)$ и $(B - \lambda I)$ были эквивалентны.

По этой причине инвариантные множители $(A - \lambda I)$ иногда называют инвариантами подобия A .

Снова из § 14 следует, что если мы напомним

$$E_i(\lambda) = (\lambda_1 - \lambda)^{i_1} (\lambda_2 - \lambda)^{i_2} \dots (\lambda_s - \lambda)^{i_s}, \quad (20.6)$$

где $\lambda_1, \lambda_2, \dots, \lambda_s$ — это различные собственные значения A , то множители в правой части (20.6) — элементарные делители A .

Треугольная каноническая форма

21. Приведение к жордановой канонической форме требует полной информации о всех собственных значениях и собственных векторах. Эта форма является частным случаем *треугольной канонической формы*. Последняя значительно более проста как с теоретической точки зрения, так и с точки зрения вычислительной практики. Ее важность вытекает из следующей теоремы и из теоремы § 25.

Каждая матрица подобна *треугольной* матрице, т. е. матрице, у которой все элементы над диагональю равны нулю. Мы будем называть такие

матрицы *нижними треугольными матрицами*. Как и в случае жордановой формы, существует также преобразование, приводящее к матрице с нулями под диагональю (*верхней треугольной*). Доказательство, сравнительно элементарное, мы отложим до § 42, где будут введены соответствующие матрицы преобразования.

Очевидно, что диагональные элементы треугольной матрицы суть собственные значения с соответствующими кратностями подобной ей исходной матрицы. Действительно, если это элементы μ_i , то характеристический полином треугольной матрицы равен $\prod (\mu_i - \lambda)$, а мы видели, что характеристический полином инвариантен при преобразованиях подобия.

Эрмитовы и симметричные матрицы

22. Как мы увидим в § 31, симметричные матрицы с вещественными элементами играют особо важную роль на практике. Мы могли бы расширить этот класс, включив в него все симметричные матрицы с комплексными элементами. Однако можно доказать, что такое расширение мало ценно, так как матрицы такого расширенного класса не обладают многими важными свойствами вещественных симметричных матриц. Наиболее используемым расширением является расширение до класса матриц вида $(P + iQ)$, где P — вещественная и симметричная, а Q — вещественная и *кососимметричная*, т. е.

$$Q^T = -Q. \quad (22.1)$$

Матрицы этого класса называются *эрмитовыми матрицами*. Если A — эрмитова матрица, то мы имеем по определению

$$\bar{A}^T = A, \quad (22.2)$$

где $\bar{A}$ обозначает матрицу с элементами, комплексно сопряженными элементам A . Матрица $\bar{A}^T$ часто обозначается A^H и называется *эрмитово сопряженной* или просто *сопряженной* матрицей. Аналогично x^H — это вектор-строка с элементами, комплексно сопряженными элементам вектора-столбца x . Такое обозначение удобно тем, что результаты, доказанные для эрмитовых матриц, могут быть переформулированы в соответствующие результаты для вещественных симметричных матриц заменой A^H на A^T и x^H на x^T . Если a — скаляр, a^H по определению есть комплексно сопряженное с a число. Из определения немедленно следует, что

$$\left. \begin{aligned} (A^H)^H &= A, \\ (x^H)^H &= x, \\ (ABC)^H &= C^H B^H A^H; \end{aligned} \right\} \quad (22.3)$$

и

$$A^H = A, \quad (22.4)$$

если A — эрмитова.

Заметим, что $y^H x$ есть скалярное произведение y и x в обычном смысле и

$$\overline{y^H x} = (y^H x)^H = x^H y. \quad (22.5)$$

Скалярное произведение $x^H x$ вещественно и положительно для всех x , отличных от нулевого вектора, так как оно равно сумме квадратов модулей компонент вектора x .

Элементарные свойства эрмитовых матриц

23. Все собственные значения эрмитовых матриц вещественны. Действительно, если

$$Ax = \lambda x, \quad (23.1)$$

то

$$x^H Ax = \lambda x^H x. \quad (23.2)$$

Число $x^H x$ вещественно и положительно для ненулевых x . Далее мы имеем

$$(x^H Ax)^H = x^H A^H x^{HH} = x^H Ax, \quad (23.3)$$

и так как $x^H Ax$ — скаляр, это значит, что он веществен.

Следовательно, λ вещественно. Для вещественных симметричных матриц вещественность собственных значений означает и вещественность собственных векторов, но собственные векторы комплексных эрмитовых матриц, вообще говоря, комплексны.

Если $Ax = \lambda x$, то для эрмитовой матрицы

$$A^H x = \lambda x, \quad (23.4)$$

и переходя к комплексному сопряжению в обеих частях, получим

$$A^T \bar{x} = \lambda \bar{x}. \quad (23.5)$$

Собственные векторы транспонированной эрмитовой матрицы, следовательно, равны комплексно сопряженным собственным векторам самой матрицы. В обозначениях § 3

$$y_i = \bar{x}_i. \quad (23.6)$$

Величины $y_i^T x_i$, играющие важную роль в теории, для эрмитовых матриц становятся $x_i^H x_i$. Из общей теории § 3 немедленно следует, что если эрмитова матрица имеет различные собственные значения, то ее собственные векторы удовлетворяют соотношениям

$$x_i^H x_j = 0 \quad (i \neq j). \quad (23.7)$$

Если мы нормируем x_i так, что

$$x_i^H x_i = 1, \quad (23.8)$$

то из уравнений (23.7) и (23.8) следует, что матрица X , образованная из собственных векторов, удовлетворяет соотношению

$$X^H X = I, \quad \text{т. е.} \quad X^H = X^{-1}. \quad (23.9)$$

Матрицы, удовлетворяющие уравнениям (23.9), называются *унитарными* матрицами. Вещественные унитарные матрицы называются *ортогональными* матрицами.

Итак, доказано для эрмитовых и вещественных симметричных матриц с различными собственными значениями следующее.

(I) Если A — эрмитова, то существует унитарная матрица U такая, что $U^H A U = \text{diag}(\lambda_i)$ (λ_i — вещественные).

(II) Если A — вещественная симметричная, то существует ортогональная матрица U такая, что $U^T A U = \text{diag}(\lambda_i)$ (λ_i — вещественные).

Может быть наиболее важным свойством эрмитовых (вещественных симметричных) матриц является то, что свойства (I), (II), которые мы пока что доказали для случая различных собственных значений, остаются справедливыми и в случае кратных собственных значений. Доказательство этого отложим до § 47.

Из этого основного результата немедленно вытекают следствия:

- (i) все элементарные делители эрмитовой матрицы линейны;
- (ii) если некоторые собственные значения эрмитовой матрицы кратны, то она неполная;
- (iii) эрмитова матрица не может быть недостаточной.

Комплексные симметричные матрицы

24. Многие важнейшие свойства вещественных симметричных матриц не распространяются на комплексные симметричные матрицы, что можно увидеть, рассмотрев матрицу A вида

$$A = \begin{bmatrix} 2i & 1 \\ 1 & 0 \end{bmatrix}. \quad (24.1)$$

Ее характеристическое уравнение есть $\lambda^2 - 2i\lambda - 1 = 0$, так что $\lambda = i$ — двукратное собственное значение. Компоненты соответствующего собственного вектора должны удовлетворять системе

$$ix_1 + x_2 = 0, \quad x_1 - ix_2 = 0, \quad (24.2)$$

так что есть лишь один собственный вектор с компонентами $1, -i$.

Мы видим, что собственному значению $\lambda = i$ соответствует квадратичный делитель. На самом деле, жорданова каноническая форма такова:

$$\begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}, \quad (24.3)$$

и мы имеем

$$\begin{bmatrix} 2i & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -i & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -i & 1 \end{bmatrix} \begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}. \quad (24.4)$$

Приведение к треугольной форме унитарными преобразованиями

25. Так как эрмитовы матрицы всегда можно привести унитарным преобразованием подобия к диагональному виду, интересно рассмотреть вопрос о том, что можно сделать такими преобразованиями для матриц общего вида. Можно показать, что любая матрица может быть приведена унитарным преобразованием к треугольному виду. Так как собственные значения треугольной матрицы определяются сразу, а ее собственные векторы можно найти сравнительно просто, этот результат имеет практическую важность, ибо унитарные преобразования обладают рядом желаемых численных свойств. Отложим доказательство до § 47, где будут введены соответствующие матрицы преобразования.

Квадратичные формы

26. Каждой вещественной симметричной матрице A мы можем сопоставить *квадратичную форму* $x^T A x$ от n переменных $x_1, x_2, \dots, x_n$

$$x^T A x = \sum_{i,j=1}^n a_{ij} x_i x_j, \quad (26.1)$$

где x — это вектор с компонентами x_i . Соответственно каждой квадратичной форме можно сопоставить *билинейную форму* $x^T A y$:

$$x^T A y = \sum_{i,j=1}^n a_{ij} x_i y_j = y^T A x. \quad (26.2)$$

Некоторые типы квадратичных форм имеют практический интерес. Квадратичная форма с матрицей A называется:

$$\left. \begin{array}{l} \text{положительно определенной,} \\ \text{отрицательно определенной,} \\ \text{неотрицательной,} \\ \text{неположительной,} \end{array} \right\} \text{ если } x^T A x \left\{ \begin{array}{l} > 0, \\ < 0, \\ \geq 0, \\ \leq 0, \end{array} \right. \text{ (для всех вещественных } x \neq 0 \text{)}.$$

Матрица положительно определенной квадратичной формы называется положительно определенной матрицей.

Необходимым и достаточным условием положительной определенности $x^T A x$ является положительность всех собственных значений A . Действительно, мы знаем, что существует ортогональная матрица R такая, что

$$R^T A R = \text{diag} (\lambda_i). \quad (26.3)$$

Следовательно, если

$$x = R z, \quad \text{т. е.} \quad z = R^T x, \quad (26.4)$$

то

$$x^T A x = z^T R^T A R z = \sum_{i=1}^n \lambda_i z_i^2. \quad (26.5)$$

Необходимым и достаточным условием положительности $\sum_{i=1}^n \lambda_i z_i^2$ для всех ненулевых z является положительность всех λ_i . Из (26.4) мы видим, что ненулевым z соответствуют ненулевые x , и наоборот; следовательно, результат установлен. Так как определитель матрицы равен произведению ее собственных значений, определитель положительно определенной матрицы должен быть положительным.

Из уравнений (26.4) и (26.5) мы видим, что проблема нахождения ортогональной матрицы R и, следовательно, собственных векторов A совпадает с проблемой нахождения главных осей поверхности второго порядка

$$\sum_{i,j} a_{ij} x_i x_j = 1 \quad (26.6)$$

и что λ_i обратны квадратам главных осей. Кратные собственные значения связываются с неопределенностью главных осей. В качестве типичного примера мы можем взять трехмерный эллипсоид с двумя равными главными осями. Одно из главных сечений в этом случае круг, и мы можем взять любые два его перпендикулярных диаметра в качестве главных осей.

Необходимые и достаточные условия положительной определенности

27. Обозначим через $x^T A_r x$ квадратичную форму, которая получится из $x^T A x$, если положить $x_{r+1}, x_{r+2}, \dots, x_n$ равными нулю. Тогда $x^T A_r x$, очевидно, — квадратичная форма r переменных $x_1, x_2, \dots, x_r$, и матрица A_r этой формы — это ведущая главная подматрица порядка r матри-

цы A . Ясно, что если $x^T A x$ — положительно определенная, то и $x^T A_r x$ должна быть положительно определенной. Действительно, если существует ненулевая последовательность значений $x_1, x_2, \dots, x_r$, для которой $x^T A_r x$ неположительна, то $x^T A x$ неположительна для вектора x ,

$$x^T = (x_1, x_2, \dots, x_r, 0, \dots, 0). \quad (27.1)$$

Следовательно, все ведущие главные миноры положительно определенной матрицы должны быть положительны.

Заметим, что точно таким же методом доказывается, что любая главная подматрица положительно определенной матрицы положительно определена и, следовательно, имеет положительный определитель. Однако наличие n положительных главных ведущих миноров является также и *достаточным* условием положительной определенности. Доказательство проводится по индукции.

Пусть утверждение справедливо для матриц $(n-1)$ порядка. Покажем, что оно верно для матриц порядка n . Предположим противное, что A — матрица порядка n , у которой n ведущих главных миноров положительны, причем A не положительно определенная. Если R — ортогональная матрица, для которой $R^T A R = \text{diag}(\lambda_i)$, то, положив $x = Rz$, получим

$$x^T A x = z^T R^T A R z = \sum \lambda_i z_i^2. \quad (27.2)$$

По предположению, A не положительно определенная, и, следовательно, так как $\prod_{i=1}^n (\lambda_i) = \det(A) > 0$, должно существовать четное число r отрицательных λ_i . Предположим, что отрицательны первые r собственных значений. Рассмотрим систему уравнений

$$z_{r+1} = z_{r+2} = \dots = z_n = x_n = 0. \quad (27.3)$$

Так как z_i — линейные функции от x_i , имеем $(n-r+1)$ линейных однородных уравнений с n неизвестными $x_1, x_2, \dots, x_n$. Поэтому существует по крайней мере одно ненулевое решение x , которое мы можем обозначить $(x_1, x_2, \dots, x_{n-1}, 0)$. Соответствующее z также ненулевое, и мы его обозначим $(z_1, z_2, \dots, z_r, 0, \dots, 0)$. Для этих x и z имеем

$$x^T A x = \sum_1^n \lambda_i z_i^2 = \sum_1^r \lambda_i z_i^2 < 0. \quad (27.4)$$

Теперь так как $x_n = 0$, это означает, что $x^T A_{n-1} x$ отрицательна для вектора $(x_1, x_2, \dots, x_{n-1})$. Это противоречит нашей гипотезе. Наш результат, очевидно, справедлив для матриц первого порядка и, следовательно, для матриц любого порядка.

Совершенно таким же способом, каким вещественной симметричной матрице мы сопоставили квадратичную форму, можно эрмитовой матрице сопоставить *эрмитову форму*:

$$x^H A x = \sum_{i,j=1}^n a_{ij} x_i \bar{x}_j. \quad (27.5)$$

Имеем

$$(x^H A x)^H = x^H A^H x^{HH} = x^H A x, \quad (27.6)$$

и, следовательно, $x^H A x$ вещественна. Концепция положительной определенности и другие подобные понятия легко переносятся на эрмитовы формы, и для них справедливы результаты, аналогичные доказанным для вещественных квадратичных форм.

Дифференциальные уравнения с постоянными коэффициентами

28. Решение алгебраической проблемы собственных значений тесно связано с решением системы линейных обыкновенных дифференциальных уравнений с постоянными коэффициентами. Общая система однородных уравнений первого порядка с n неизвестными $y_1, y_2, \dots, y_n$ может быть записана в виде

$$B \cdot \frac{d}{dt}(y) = Cy, \quad (28.1)$$

где t — независимая переменная, y — вектор с компонентами y_i , а B и C — матрицы n -го порядка. Если B — особенная, то левые части системы уравнений связаны линейным соотношением. Правые части должны удовлетворять тому же самому соотношению, и, следовательно, y_i линейно зависимы. Систему уравнений можно поэтому привести к системе низшего порядка. В дальнейшем будем полагать, что B — неособенная.

Тогда уравнение (28.1) можно записать в виде

$$\frac{d}{dt}(y) = Ay, \quad (28.2)$$

где

$$A = B^{-1}C. \quad (28.3)$$

Предположим, что (28.2) имеет решение вида

$$y = x e^{\lambda t}, \quad (28.4)$$

где x — это вектор, не зависящий от t . Тогда

$$\lambda x e^{\lambda t} = A x e^{\lambda t},$$

и поэтому

$$\lambda x = A x. \quad (28.5)$$

Следовательно, уравнение (28.4) будет давать решение уравнения (28.2), если λ — собственное значение A , а x — соответствующий ему собственный вектор. Если у A имеется r линейно независимых собственных векторов, то мы получим r линейно независимых решений (28.2), независимо от того, будут ли различны соответствующие собственные значения. Однако уравнение (28.2) должно иметь n линейно независимых решений, так что если $r < n$, т. е. если некоторые элементарные делители A нелинейны, то (28.2) будет иметь решения не чисто экспоненциального типа.

Решения, соответствующие нелинейным элементарным делителям

29. Природу решений в случае, когда есть нелинейные элементарные делители, можно изучить в терминах жордановой канонической формы. Пусть X — матрица подобного преобразования, приводящего A к жордановой форме. Если мы введем новые переменные z , положив

$$y = Xz, \quad (29.1)$$

то (28.2) перейдет в

$$X \frac{d}{dt}(z) = AXz, \quad (29.2)$$

или

$$\frac{d}{dt}(z) = X^{-1}AXz. \quad (29.3)$$

Заметим, что матрица системы при этом претерпела преобразование подобия и превратилась в жорданову каноническую матрицу.

Предположим, например, что A — матрица шестого порядка с нижней жордановой канонической формой

$$\begin{bmatrix} C_3(\lambda_1) & & \\ & C_2(\lambda_1) & \\ & & C_1(\lambda_2) \end{bmatrix}.$$

Тогда уравнения (29.3) суть

$$\left[\begin{array}{l} \frac{dz_1}{dt} = \lambda_1 z_1 \\ \frac{dz_2}{dt} = z_1 + \lambda_1 z_2 \\ \frac{dz_3}{dt} = z_2 + \lambda_1 z_3 \\ \frac{dz_4}{dt} = \lambda_1 z_4 \\ \frac{dz_5}{dt} = z_4 + \lambda_1 z_5 \\ \frac{dz_6}{dt} = \lambda_2 z_6 \end{array} \right]. \quad (29.4)$$

Решениями первого, четвертого и шестого из этих уравнений будут

$$z_1 = a_1 e^{\lambda_1 t}, \quad z_4 = a_4 e^{\lambda_1 t}, \quad z_6 = a_6 e^{\lambda_2 t}, \quad (29.5)$$

где a_1 , a_4 и a_6 — произвольные постоянные. Общее решение второго уравнения имеет вид

$$z_2 = a_2 e^{\lambda_1 t} + a_1 t e^{\lambda_1 t}, \quad (29.6)$$

третьего —

$$z_3 = a_3 e^{\lambda_1 t} + a_2 t e^{\lambda_1 t} + \frac{1}{2} a_1 t^2 e^{\lambda_1 t} \quad (29.7)$$

и пятого —

$$z_5 = a_5 e^{\lambda_1 t} + a_4 t e^{\lambda_1 t}, \quad (29.8)$$

где a_2 , a_3 и a_5 — произвольные постоянные. Общим решением системы, следовательно, будет

$$\left. \begin{aligned} z_1 &= a_1 e^{\lambda_1 t}, \\ z_2 &= a_1 t e^{\lambda_1 t} + a_2 e^{\lambda_1 t}, \\ z_3 &= \frac{1}{2} a_1 t^2 e^{\lambda_1 t} + a_2 t e^{\lambda_1 t} + a_3 e^{\lambda_1 t}, \\ z_4 &= a_4 e^{\lambda_1 t}, \\ z_5 &= a_4 t e^{\lambda_1 t} + a_5 e^{\lambda_1 t}, \\ z_6 &= a_6 e^{\lambda_2 t}. \end{aligned} \right\} \quad (29.9)$$

Заметим, что лишь три линейно независимых решения имеют чисто экспоненциальный вид. Именно:

$$z = a_3 e^{\lambda_1 t} e_3, \quad z = a_5 e^{\lambda_1 t} e_5, \quad z = a_6 e^{\lambda_2 t} e_6, \quad (29.10)$$

как мы и могли ожидать, так как e_3 , e_5 и e_6 суть все собственные векторы жордановой канонической матрицы.

Общее решение исходной системы уравнений получается подстановкой z в (29.1). Если столбцы X обозначить $x_1, x_2, \dots, x_6$, то общее решение (28.2) имеет вид

$$\begin{aligned} y &= a_1 e^{\lambda_1 t} \left(x_1 + t x_2 + \frac{1}{2} t^2 x_3 \right) + a_2 e^{\lambda_1 t} (x_2 + t x_3) + \\ &\quad + a_3 e^{\lambda_1 t} x_3 + a_4 e^{\lambda_1 t} (x_4 + t x_5) + a_5 e^{\lambda_1 t} x_5 + a_6 e^{\lambda_2 t} x_6. \end{aligned} \quad (29.11)$$

Результат в общем случае совершенно подобен этому.

Дифференциальные уравнения высшего порядка

30. Рассмотрим систему n однородных дифференциальных уравнений r -го порядка вида

$$A_r \frac{d^r}{dt^r} (y_0) + A_{r-1} \frac{d^{r-1}}{dt^{r-1}} (y_0) + \dots + A_0 y_0 = 0, \quad (30.1)$$

где $A_r, A_{r-1}, \dots, A_0$ — матрицы n -го порядка с постоянными элементами, $\det(A_r) \neq 0$, а y_0 есть n -мерный вектор, образованный n неизвестными функциями. (Мы обозначаем y_0 вместо y по причинам, которые скоро станут понятными.) Систему (30.1) можно свести к системе уравнений первого порядка введением $n(r-1)$ новых переменных

$$\left. \begin{aligned} y_1 &= \frac{d}{dt} (y_0), \\ y_2 &= \frac{d}{dt} (y_1), \\ &\dots \dots \dots \\ y_{r-1} &= \frac{d}{dt} (y_{r-2}). \end{aligned} \right\} \quad (30.2)$$

Если A_r — неособенная, мы можем записать (30.1) в виде

$$B_0 y_0 + B_1 y_1 + \dots + B_{r-1} y_{r-1} = \frac{d}{dt} (y_{r-1}), \quad (30.3)$$

где

$$B_s = -A_r^{-1}A_s. \quad (30.4)$$

Уравнения (30.2) и (30.3) можно рассматривать как одну систему nr уравнений с nr переменными, которые являются компонентами $y_0, y_1, \dots, y_{r-1}$. Мы можем записать ее в форме

$$\begin{bmatrix} 0 & I & & 0 \\ 0 & 0 & I & 0 \\ 0 & & 0 & I \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & I \\ B_0 & B_1 & B_2 & B_3 & \dots & B_{r-1} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{r-2} \\ y_{r-1} \end{bmatrix} = \frac{d}{dt} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{r-2} \\ y_{r-1} \end{bmatrix}. \quad (30.5)$$

Эту систему можно решить методом предыдущего параграфа, что требует определения жордановой канонической формы матрицы B nr -го порядка, стоящей в левой части (30.5). Хотя эта матрица высокого порядка, она имеет весьма простую форму и большое количество нулевых элементов. Если λ — собственное значение B и x — собственный вектор, который разделен на r векторов $x_0, x_1, \dots, x_{r-1}$, то

$$\left. \begin{aligned} x_i &= \lambda x_{i-1} \quad (i=1, \dots, r-1), \\ B_0 x_0 + B_1 x_1 + \dots + B_{r-1} x_{r-1} &= \lambda x_{r-1} \end{aligned} \right\} \quad (30.6)$$

и, следовательно,

$$(B_0 + B_1 \lambda + \dots + B_{r-1} \lambda^{r-1}) x_0 = \lambda^r x_0. \quad (30.7)$$

Эти уравнения имеют ненулевое решение, если

$$\det(B_0 + B_1 \lambda + \dots + B_{r-1} \lambda^{r-1} - I \lambda^r) = 0. \quad (30.8)$$

На практике мы имеем выбор — работать с матрицей B порядка nr и решать стандартную задачу на собственные значения или же иметь дело с уравнением (30.8), которое нестандартно, но в которое входят матрицы лишь порядка n .

Уравнения второго порядка специального вида

31. Уравнение малых колебаний механической системы около положения устойчивого равновесия под воздействием консервативных сил имеет вид

$$B\ddot{y} = -Ay, \quad (31.1)$$

где A и B симметричны и положительно определены. Если мы предположим, что решения имеют вид $y = xe^{i\mu t}$, то мы должны иметь

$$\mu^2 Bx = Ax. \quad (31.2)$$

Положив $\mu^2 = \lambda$, получим

$$\lambda Bx = Ax. \quad (31.3)$$

Мы сейчас покажем, что если A и B симметричны и B — положительно определенная, то все корни $\det(A - \lambda B)$ вещественны.

Так как B — положительно определенная, то существует ортогональная матрица R такая, что

$$R^T B R = \text{diag}(\beta_i^2), \quad (31.4)$$

где β_i^2 , собственные значения B , обязательно положительные. Можно написать

$$\text{diag}(\beta_i) = D, \quad \text{diag}(\beta_i^2) = D^2. \quad (31.5)$$

Тогда мы имеем

$$(A - \lambda B) = RD(D^{-1}R^TARD^{-1} - \lambda I)DR^T, \quad (31.6)$$

и, следовательно,

$$\det(A - \lambda B) = (\det R)^2 (\det D)^2 \det(P - \lambda I), \quad (31.7)$$

где

$$P = D^{-1}R^TARD^{-1}. \quad (31.8)$$

Так как $(\det R)^2 = 1$ и $(\det D)^2 = \prod \beta_i^2$, нули $\det(A - \lambda B)$ совпадают с нулями $\det(P - \lambda I)$. Последние же суть собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$ симметричной матрицы P и поэтому вещественны.

Матрица P имеет полный набор собственных векторов z_i , которые можно выбрать ортогональными. Поэтому имеем

$$Pz_i = \lambda_i z_i, \quad D^{-1}R^TARD^{-1}z_i = \lambda_i z_i, \quad (31.9)$$

что дает

$$A(RD^{-1}z_i) = \lambda_i RDz_i = \lambda_i RD(DR^T RD^{-1})z_i = \lambda_i B(RD^{-1}z_i). \quad (31.10)$$

Следовательно, $x_i = RD^{-1}z_i$ это «собственный вектор», соответствующий λ_i в обобщенной задаче на собственные значения, определяемой уравнением (31.3). Заметим, что так как z_i ортогональны, а RD^{-1} — вещественная и неособенная, x_i образуют полную систему вещественных собственных векторов уравнения (31.3). Далее имеем

$$0 = z_i^T z_j = (DR^T x_i)^T DR^T x_j = x_i^T RDDR^T x_j = x_i^T Bx_j, \quad (31.11)$$

что показывает, что векторы x_i ортогональны относительно B .

Если A — тоже положительно определенная, то λ_i положительны. Действительно, мы имеем

$$Ax_i = \lambda_i Bx_i, \quad x_i^T Ax_i = \lambda_i x_i^T Bx_i, \quad (31.12)$$

причем $x_i^T Ax_i$ и $x_i^T Bx_i$ положительны.

Явное решение уравнения $B\ddot{y} = -Ay$

32. Мы можем теперь вывести явное выражение для решения (31.1), когда A и B — положительно определенные. Пусть $x_1, x_2, \dots, x_n$ — n линейно независимых решений (31.3), соответствующих собственным значениям $\lambda_1, \lambda_2, \dots, \lambda_n$. Последние должны быть положительными, но не обязательно различными. Положим

$$\lambda_i = \mu_i^2, \quad (32.1)$$

где μ_i выбраны положительными. Тогда система уравнений (31.1) имеет общее решение

$$y = \sum_1^n (a_i e^{i\mu_i t} + b_i e^{-i\mu_i t}) x_i, \quad (32.2)$$

с $2n$ произвольными постоянными a_i и b_i . Предполагая, что решение вещественно, мы можем записать его в виде

$$y = \sum_1^n c_i \cos(\mu_i t + \varepsilon_i) x_i, \quad (32.3)$$

где c_i и ε_i — произвольные постоянные.

Уравнения вида $(AB - \lambda I)x = 0$

33. В теоретической физике проблема, которую мы сейчас рассмотрели, часто возникает в форме

$$(AB - \lambda I)x = 0, \quad (33.1)$$

где A и B — симметричные, и одна из них (или обе) — положительно определенная. Если A — положительно определенная, то (33.1) эквивалентно

$$(B - \lambda A^{-1})x = 0, \quad (33.2)$$

так как положительно определенная матрица имеет положительный определитель и, следовательно, неособенная. Если B — положительно определенная, то (33.1) можно записать в виде

$$(A - \lambda B^{-1})(Bx) = 0. \quad (33.3)$$

Заметим, что матрица, обратная для положительно определенной матрицы, положительно определенная. Действительно, если

$$A = R^T \text{diag}(\alpha_i^2) R, \quad (33.4)$$

то

$$A^{-1} = R^T \text{diag}(\alpha_i^{-2}) R. \quad (33.5)$$

Следовательно, уравнение (33.1) не отличается существенно от уравнения (31.3).

Важность положительной определенности одной из матриц не должна недооцениваться. Если A и B — вещественные симметричные матрицы, но ни одна из них не положительно определенная, то собственные значения AB не обязаны быть вещественными, и это же справедливо для корней $\det(A - \lambda B) = 0$. Простой пример показывает это. Пусть

$$A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (33.6)$$

Тогда

$$AB = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}. \quad (33.7)$$

Собственные значения AB удовлетворяют уравнению

$$\lambda^2 = ab \quad (33.8)$$

и они комплексны, если a и b различных знаков.

Если ни A , ни B не положительно определенные, и $\alpha + \beta i$ — комплексный корень $\det(A - \lambda B) = 0$, то мы имеем

$$Ax = (\alpha + \beta i)Bx \quad (33.9)$$

для некоторого ненулевого x . Следовательно,

$$x^H A x = (\alpha + \beta i) x^H B x. \quad (33.10)$$

Так как $x^H A x$ и $x^H B x$ — вещественные, уравнение (33.10) означает, что

$$x^H A x = x^H B x = 0. \quad (33.11)$$

Минимальный полином вектора

34. Собственный вектор x матрицы характеризуется тем, что существует линейная зависимость между x и Ax . Для произвольного вектора b такой зависимости не существует, но если мы образуем последовательность $b, Ab, A^2b, \dots, A^r b$, то должно существовать наименьшее значение m для r , при котором эти векторы линейно зависимы, и ясно, что $m \leq n$. Мы можем записать соответствующее соотношение в виде

$$(A^m + c_{m-1}A^{m-1} + \dots + c_0I)b = 0. \quad (34.1)$$

По определению m не существует соотношений вида

$$(A^r + d_{r-1}A^{r-1} + \dots + d_0I)b = 0 \quad (r < m). \quad (34.2)$$

Нормализованный полином $s(\lambda)$, соответствующий левой части (34.1), называется *минимальным* *) *полиномом вектора b относительно A* . Если $s(A)$ — любой другой полином от A , для которого $s(A)b = 0$, то $s(A)$ должен делиться на $c(A)$. Действительно, если это не так, то по алгоритму Евклида существуют полиномы $p(A)$, $q(A)$, $r(A)$ такие, что

$$p(A)s(A) - q(A)c(A) = r(A), \quad (34.3)$$

где $r(A)$ низшей степени, чем $c(A)$. Но из (34.3) следует, что $r(A)b = 0$, и это противоречит гипотезе, что $c(A)$ — полином наименьшей степени, для которого это верно.

Минимальный полином единствен, ибо если $d(A)b = 0$ для второго нормализованного полинома степени m , то $[c(A) - d(A)]b = 0$, и $c(A) - d(A)$ степени меньшей, чем m . Степень $c(A)$ называется *высотой b относительно A* .

Минимальный полином матрицы

35. Мы можем подобным образом рассмотреть последовательность матриц $I, A, A^2, \dots, A^r$. Снова существует наименьшее значение s для r , при котором эта последовательность матриц линейно зависима. Мы можем записать это в виде соотношения

$$A^s + m_{s-1}A^{s-1} + \dots + m_0I = 0. \quad (35.1)$$

Например, если

$$A = \begin{bmatrix} p_1 & p_0 \\ 1 & 0 \end{bmatrix}, \quad (35.2)$$

то

$$A^2 = \begin{bmatrix} p_1^2 + p_0 & p_1 p_0 \\ p_1 & p_0 \end{bmatrix}, \quad (35.3)$$

*) Иногда такой полином называют минимальным, аннулирующим вектор b , полиномом. (Прим. перев.)

и, следовательно,

$$A^2 - p_1 A - p_0 I = \begin{bmatrix} p_1^2 + p_0 & p_0 p_1 \\ p_1 & p_0 \end{bmatrix} - \begin{bmatrix} p_1^2 & p_1 p_0 \\ p_1 & 0 \end{bmatrix} - \begin{bmatrix} p_0 & 0 \\ 0 & p_0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (35.4)$$

Соотношения более низкой степени не существует, так как

$$A + c_0 I = \begin{bmatrix} p_1 + c_0 & p_0 \\ 1 & c_0 \end{bmatrix} \quad (35.5)$$

и это не равно нулю ни при каком c_0 . Этот пример важен для следующих параграфов.

Полином $m(\lambda)$, соответствующий левой части уравнения (35.4), называется *минимальным полиномом матрицы* A . Точно так же, как в § 34, можем показать, что он единствен и делит все другие полиномы $p(\lambda)$, для которых $p(A) = 0$.

Минимальный полином $c(\lambda)$ любого вектора b по отношению к A делит $m(\lambda)$, так как мы, очевидно, имеем $m(A)b = 0$. Пусть $c_1(\lambda)$, $c_2(\lambda)$, ..., $c_n(\lambda)$ — минимальные полиномы e_1 , e_2 , ..., e_n по отношению к A , и пусть $g(\lambda)$ — наименьшее общее кратное (н. о. к.) $c_i(\lambda)$. Тогда имеем

$$g(A)e_i = 0 \quad (i = 1, \dots, n), \quad (35.6)$$

и, следовательно, $g(A) = 0$. Отсюда следует, что $g(A)b = 0$ для любого вектора, и поэтому $g(\lambda)$ это н. о. к. минимальных полиномов всех векторов.

На самом деле мы должны иметь $g(\lambda) = m(\lambda)$, так как по определению $m(A) = 0$, и, следовательно, $m(A)e_i = 0$. Наименьшее общее кратное минимальных полиномов e_i должно поэтому делить $m(\lambda)$, и, следовательно, $g(\lambda)$ делит $m(\lambda)$. Но по определению $m(\lambda)$ — это полином наименьшей степени, для которого $m(A) = 0$.

Мы развили эту теорию, исходя из системы векторов-столбцов $b, Ab, A^2b, \dots$. Аналогично, рассматривая системы векторов-строк $b^T, b^T A, b^T A^2, \dots$, мы можем определить минимальный полином b^T по отношению к A . Ясно, что минимальный полином b^T по отношению к A тот же самый, что и минимальный полином b по отношению к A^T .

Если $h(\lambda)$ это н. о. к. минимальных полиномов $e_1^T, e_2^T, \dots, e_n^T$, то мы можем показать, что $h(A)$ — нулевая матрица, и следовательно, $h(\lambda)$ тоже совпадает с $m(\lambda)$. Рассматривая векторы-строки или векторы-столбцы, мы получаем тот же минимальный полином A . Нам будет в дальнейшем (§ 37) удобнее работать с векторами-строками.

Теорема Кэли — Гамильтона

36. Так как A содержит n^2 элементов, то ясно, что минимальный полином A не может быть степени выше чем n^2 . Сейчас покажем, что если характеристическое уравнение A имеет вид

$$\lambda^n + c_{n-1}\lambda^{n-1} + \dots + c_0 = 0, \quad (36.1)$$

то

$$A^n + c_{n-1}A^{n-1} + \dots + c_0 I = 0. \quad (36.2)$$

Это означает, что минимальный полином матрицы не может быть степени выше чем n .

где $f(A)$ — любой полином от A . Если $f(A) = 0$, то и $f(B) = 0$, и наоборот. Следовательно, минимальный полином B делится на минимальный полином A , и наоборот. Такие полиномы должны совпадать.

Рассмотрим матрицу Фробениуса B_r порядка r , заданную уравнением (13.1). Имеем

$$\left. \begin{aligned} e_{s+1}^T B_r &= e_s^T & (s=1, 2, \dots, r-1), \\ e_1^T B_r &= (b_{r-1}, b_{r-2}, \dots, b_0), \end{aligned} \right\} \quad (37.3)$$

что дает

$$e_r^T (B_r)^k = e_{r-k}^T \quad (k=0, 1, \dots, r-1). \quad (37.4)$$

Поэтому мы имеем

$$e_r^T (\alpha_k B_r^k + \dots + \alpha_0 I) = (0, 0, \dots, 0, \alpha_k, \alpha_{k-1}, \dots, \alpha_0), \quad k \leq r-1, \quad (37.5)$$

и вектор справа нулевой, только если все α_i равны нулю. Поэтому степень минимального полинома e_r^T относительно B_r не меньше чем r . Более того,

$$e_r^T B_r^r = e_r^T (b_{r-1} B_r^{r-1} + b_{r-2} B_r^{r-2} + \dots + b_0 I), \quad (37.6)$$

и, следовательно, минимальный полином e_r^T это характеристический полином B_r . Поэтому минимальный полином B_r является также ее характеристическим полиномом. Пример § 35 дает иллюстрацию этого в случае $r = 2$.

Возвращаясь к общей канонической форме Фробениуса B , мы можем обозначить ее

$$B = \begin{bmatrix} B_{r_1} & & \\ & B_{r_2} & \\ & & \ddots \\ & & & B_{r_s} \end{bmatrix}, \quad (37.7)$$

где характеристический полином каждой B_{r_i} делится на характеристический полином последующей. Для любого полинома $f(B)$ имеем

$$f(B) = \begin{bmatrix} f(B_{r_1}) & & \\ & f(B_{r_2}) & \\ & & \ddots \\ & & & f(B_{r_s}) \end{bmatrix}. \quad (37.8)$$

Если $f(B)$ — нулевая матрица, то каждая матрица $f(B_{r_i})$ должна быть нулевой. Так как $f(B_{r_i})$ будет нулевой матрицей тогда и только тогда, когда $f(\lambda)$ кратен $f_i(\lambda)$ — характеристическому полиному B_{r_i} , и так как каждый $f_i(\lambda)$ делит $f_1(\lambda)$, то $f_1(\lambda)$ это минимальный полином B . Следовательно, $f_1(\lambda)$ — минимальный полином любой матрицы, подобной B .

Наше доказательство показывает, что минимальный полином матрицы равен ее характеристическому полиному тогда и только тогда, когда в ее каноническую форму Фробениуса входит лишь одна матрица Фробениуса, т. е. если она полна. Это, в частности, справедливо тогда, когда все собственные значения матрицы различны.

38. Возвращаясь к жордановой канонической форме, рассмотрим сначала простую жорданову матрицу, например

$$C_3(a) = \begin{bmatrix} a & 0 & 0 \\ 1 & a & 0 \\ 0 & 1 & a \end{bmatrix}. \quad (38.1)$$

Обозначая для краткости ее через C , получим

$$C^2 = \begin{bmatrix} a^2 & 0 & 0 \\ 2a & a^2 & 0 \\ 1 & 2a & a^2 \end{bmatrix}, \quad C^3 = \begin{bmatrix} a^3 & 0 & 0 \\ 3a^2 & a^3 & 0 \\ 3a & 3a^2 & a^3 \end{bmatrix}. \quad (38.2)$$

(Здесь для определенности мы взяли нижнюю жорданову каноническую матрицу.) Очевидно, что C^2 не является линейной комбинацией C и I , и, следовательно, степень минимального полинома $C_3(a)$ больше двух. Поэтому минимальный полином — это характеристический полином $(a - \lambda)^3$. Если теперь C есть жорданова каноническая форма, то она есть прямая сумма жордановых подматриц. Обозначим эти подматрицы $C_{i_j}(\lambda_i)$, где каждому λ_i может, конечно, соответствовать несколько подматриц. Как и в случае формы Фробениуса, мы видим, что $f(C)$, полином от C , является прямой суммой матриц $f[C_{i_j}(\lambda_i)]$. Поэтому $f(C) = 0$ тогда и только тогда, когда все $f[C_{i_j}(\lambda_i)] = 0$. Если мы обозначим через $C_{i_1}(\lambda_1)$ жорданову матрицу высшего порядка, соответствующую λ_1 , то минимальный полином для C должен быть $\prod_i (\lambda - \lambda_i)^{i_1}$. Связь между каноническими формами Жордана и Фробениуса, установленная в § 14, подтверждает, что подобные канонические формы Жордана и Фробениуса имеют одинаковые минимальные полиномы.

Следовало бы проверить, что вектор, элементы которого на местах, соответствующих жордановым подматрицам высшего порядка для каждого λ_i , равны единице, а остальные равны нулю, имеет своим минимальным полиномом минимальный полином самой жордановой формы.

Корневые векторы

39. Если матрица A имеет линейные элементарные делители, то существуют n собственных векторов, на которые натянуто все n -мерное пространство. Если A имеет нелинейные делители, то это неверно, так как существуют меньше чем n линейно независимых собственных векторов. Тем не менее удобно иметь систему базисных векторов такую, чтобы она переходила в систему n собственных векторов в случае, если A имеет линейные делители. Как мы видели, в последнем случае за собственные векторы могут быть взяты столбцы матрицы X такой, что

$$X^{-1}AX = \text{diag}(\lambda_i). \quad (39.1)$$

В случае нелинейных элементарных делителей естественно взять в качестве базиса n столбцов матрицы X , приводящей A к жордановой канонической матрице.

Эти векторы удовлетворяют важным соотношениям. Достаточно продемонстрировать это на примере. Пусть матрица A 8-го порядка такая, что

$$AX = X \begin{bmatrix} C_3(\lambda_1) & & & \\ & C_2(\lambda_1) & & \\ & & C_2(\lambda_2) & \\ & & & C_1(\lambda_3) \end{bmatrix}. \quad (39.2)$$

Тогда, если $x_1, x_2, \dots, x_8$ — столбцы X , то

$$\left. \begin{aligned} Ax_1 &= \lambda_1 x_1 + x_2, & Ax_4 &= \lambda_1 x_4 + x_5, & Ax_6 &= \lambda_2 x_6 + x_7, & Ax_8 &= \lambda_3 x_8, \\ Ax_2 &= \lambda_1 x_2 + x_3, & Ax_5 &= \lambda_1 x_5, & Ax_7 &= \lambda_2 x_7, \\ Ax_3 &= \lambda_1 x_3, \end{aligned} \right\} \quad (39.3)$$

откуда выводим, что

$$\left. \begin{aligned} (A - \lambda_1 I)^3 x_1 &= 0, & (A - \lambda_1 I)^2 x_4 &= 0, & (A - \lambda_2 I)^2 x_6 &= 0, & (A - \lambda_3 I) x_8 &= 0, \\ (A - \lambda_1 I)^2 x_2 &= 0, & (A - \lambda_1 I) x_5 &= 0, & (A - \lambda_2 I) x_7 &= 0, \\ (A - \lambda_1 I) x_3 &= 0, \end{aligned} \right\} \quad (39.4)$$

Каждый из этих столбцов X , следовательно, удовлетворяет соотношению вида

$$(A - \lambda_i I)^j x_k = 0. \quad (39.5)$$

Вектор, удовлетворяющий уравнению (39.5), но не удовлетворяющий аналогичному соотношению с меньшим показателем при $(A - \lambda_i I)$, называется *корневым вектором высоты j* , соответствующим λ_i . Заметим, что $(\lambda - \lambda_i)^j$ — это минимальный полином x_k относительно A . Любой собственный вектор — это корневой вектор высоты один и столбцы матрицы X — это все корневые векторы A . Так как столбцы X линейно независимы, то существует система корневых векторов, на которые натянуто все n -мерное пространство. В общем случае базисная система корневых векторов не единственна, так как, если x — корневой вектор высоты r , соответствующий собственному значению λ_i , то это же справедливо для вектора, который получен добавлением любых корневых векторов высот, не больших r , соответствующих тому же собственному значению и умноженных на произвольные множители. Для матрицы жордановой канонической формы векторы $e_1, e_2, \dots, e_n$ образуют полную систему корневых векторов.

Элементарные преобразования подобия

40. Приведение матрицы общего вида к одной из канонических форм, или к одной из специальных форм обычно проводится последовательным применением простых преобразований подобия. Мы уже рассмотрели в § 17 матрицы некоторых элементарных преобразований, но здесь удобно несколько обобщить их и ввести стандартные обозначения, которыми мы будем пользоваться во всей книге. Мы будем называть *элементарными преобразованиями* преобразования с любой из *элементарных матриц*, которыми являются:

(i) Матрицы I_{ij} , отличающиеся от единичной лишь строками i и j и столбцами i и j , которые имеют вид

$$\begin{array}{cc} \text{столбец } i & \text{столбец } j \\ \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right] & \begin{array}{l} \text{строка } i \\ \text{строка } j. \end{array} \end{array} \quad (40.1)$$

В частности, $I_{ii} = I$. Очевидно, $I_{ij} I_{ij} = I$, так что I_{ij} является своей обратной; она также ортогональна. При умножении слева на I_{ij} переставляются i -я и j -я строки, а при умножении справа переставляются i -й и j -й столбцы. Поэтому преобразование подобия с матрицей I_{ij} переставляет i -ю и j -ю строки и i -й и j -й столбцы.

(ii) Матрицы $P_{\alpha_1, \alpha_2, \dots, \alpha_n}$, у которых

$$p_{i\alpha_i} = 1, \quad \text{остальные } p_{ij} = 0, \quad (40.2)$$

где $(\alpha_1, \alpha_2, \dots, \alpha_n)$ — некоторая перестановка $(1, 2, \dots, n)$. Эти матрицы называются *матрицами перестановок*. Заметим, что у матрицы перестановки каждая строка и каждый столбец содержат только один элемент, равный единице. Поэтому транспонированная матрица перестановки тоже является матрицей перестановки, и мы имеем

$$(P_{\alpha_1, \alpha_2, \dots, \alpha_n})^{-1} = (P_{\alpha_1, \alpha_2, \dots, \alpha_n})^T. \quad (40.3)$$

Каждая матрица перестановки может быть представлена в виде произведения матриц типа I_{ij} . Умножение слева на P переводит α_i -ю строку в i -ю строку, умножение справа на P^T переводит α_i -й столбец в i -й столбец.

(iii) Матрицы R_i , отличающиеся от единичной лишь i -й строкой, которая равна

$$(-r_{i1}, -r_{i2}, \dots, -r_{i, i-1}, 1, -r_{i, i+1}, \dots, -r_{in}). \quad (40.4)$$

При умножении A слева на R_i из i -й строки *вычитаются* остальные, умноженные на множители, причем множитель j -й строки равен r_{ij} . При умножении справа из каждого столбца вычитается i -й столбец, умноженный на соответствующий множитель. Обратная матрица R_i^{-1} — матрица того же типа, у которой каждый элемент $-r_{ij}$ заменен на $+r_{ij}$. Преобразование подобия $R_i^{-1}AR_i$ матрицы A заключается в *вычитании* i -го столбца, умноженного на соответствующие множители, из всех остальных столбцов и затем в *прибавлении* к i -й строке всех остальных, умноженных на эти же множители. Заметим, что на первом шаге меняются все столбцы, кроме i -го, на втором меняется лишь i -я строка.

(iv) Матрицы S_i , полученные транспонированием из матриц вида R_i . Столбец матрицы S_i с номером i , записанный в виде строки, есть

$$(-s_{i1}, -s_{i2}, \dots, -s_{i-1, i}, 1, -s_{i+1, i}, \dots, -s_{in}). \quad (40.5)$$

Их свойства следуют из свойств R_i с соответствующими изменениями

(v) Матрицы M_i , отличающиеся от единичной матрицы лишь i -й строкой, которая равна

$$(0, 0, \dots, 0, 1, -m_{i, i+1}, \dots, -m_{in}). \quad (40.6)$$

Очевидно, M_i это R_i , у которой $r_{i1}, \dots, r_{i, i-1}$ — нули.

(vi) Матрицы N_i , полученные транспонированием M_i . Очевидно, i -й столбец, записанный в виде вектора-строки, есть

$$(0, 0, \dots, 0, 1, -n_{i+1, i}, \dots, -n_{ni}). \quad (40.7)$$

Следовательно, N_i это S_i , у которой часть элементов — нули.

(vii) Матрицы M_{ij} , отличающиеся от единичной лишь (i, j) -ым элементом, который равен $-m_{ij}$. Таким образом, M_{ij} является частным видом ранее рассмотренных матриц.

Свойства элементарных матриц

41. Читатель может легко проверить следующие важные свойства элементарных матриц:

(i) Произведение $M_{n-1}M_{n-2} \dots M_1$ есть верхняя треугольная матрица с единичными диагональными элементами *). Элемент в (i, j) -й

*) Мы будем называть в дальнейшем такую матрицу *единичной верхней (нижней) треугольной матрицей*. (Прим. перев.)

позиции при $j > i$ равен $-m_{ij}$. Произведения элементов m_{pq} в нее не входят.

(ii) Произведение $M_1 M_2 \dots M_{n-1}$ — также верхняя треугольная матрица с единичными диагональными элементами, но произведения m_{pq} входят и в другие элементы и структура их более сложна, чем в других случаях.

(iii) Аналогично оба произведения $N_1 N_2 \dots N_{n-1}$ и $N_{n-1} \dots N_1$ это нижние треугольные матрицы с единичными диагональными элементами. У первой элемент на (i, j) -м месте при $i > j$ равен $-n_{ij}$. Последняя имеет сложную структуру, похожую на структуру $M_1 M_2 \dots M_{n-1}$.

(iv) Матрица

$$[I_{n-1, (n-1)'} \dots I_{r+2, (r+2)'} I_{r+1, (r+1)'}] M_r [I_{r+1, (r+1)'} I_{r+2, (r+2)'} \dots I_{n-1, (n-1)'}]$$

с $(r+1)' \geq r+1$, $(r+2)' \geq r+2$, ..., $(n-1)' \geq n-1$ — это матрица того же типа, что и M_r , у которой элементы в r -й строке в столбцах, начиная с $(r+1)$ -го, суть перестановка элементов M_r . Эта перестановка является результатом последовательных подстановок $[(r+1)', r+1]$, $[(r+2)', r+2]$, ..., $[(n-1)', n-1]$. Аналогичный результат остается в силе и для матриц N_r .

Приведение к треугольной канонической форме элементарными преобразованиями подобия

42. Мы можем проиллюстрировать использование элементарных преобразований подобия, показав, что любая матрица может быть приведена к треугольному виду преобразованием подобия, матрица которого является произведением элементарных матриц. Доказательство проведем по индукции.

Предположим, что это верно для матриц порядка $n-1$. Пусть A — матрица порядка n , и пусть λ — ее собственное значение. Тогда существует по крайней мере один собственный вектор x , соответствующий λ , такой, что

$$Ax = \lambda x. \quad (42.1)$$

Теперь для любой неособенной X собственный вектор $XA X^{-1}$ равен Xx . Мы покажем, как выбрать X так, чтобы Xx был равен e_1 . Пусть

$$x^T = (x_1, x_2, \dots, x_{r-1}, 1, x_{r+1}, \dots, x_n), \quad (42.2)$$

где r -й элемент — это первый ненулевой элемент x , и произвольный множитель выбран так, что он равен единице. Имеем

$$(I_{1r} x)^T = (1, y_2, y_3, \dots, y_n) = y^T, \quad (42.3)$$

где y_i — это x_i в некотором порядке. Заметим, что случай $r=1$ включен, ибо $I_{11} = I$. Теперь если мы умножим y слева на матрицу N_1 , у которой $n_{i1} = y_i$, то все элементы аннулируются, кроме первого. Следовательно,

$$N_1 y = N_1 I_{1r} x = e_1, \quad (42.4)$$

и матрица $N_1 I_{1r} A I_{1r} N_1^{-1}$ имеет собственное значение λ с соответствующим собственным вектором e_1 . Поэтому она должна иметь вид

$$\begin{bmatrix} \lambda & b^T \\ 0 & B \end{bmatrix}, \quad (42.5)$$

где B — квадратная матрица порядка $(n - 1)$. Теперь, по предположению, существует квадратная матрица H , которая является произведением элементарных матриц, такая, что

$$HBH^{-1} = T, \quad (42.6)$$

где T — треугольная. Следовательно,

$$\begin{bmatrix} 1 & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} \lambda & b^T \\ 0 & B \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & H^{-1} \end{bmatrix} = \begin{bmatrix} \lambda & b^T H^{-1} \\ 0 & HBH^{-1} \end{bmatrix} = \begin{bmatrix} \lambda & b^T H^{-1} \\ 0 & T \end{bmatrix}, \quad (42.7)$$

так что матрица в правой части треугольная. Результат поэтому верен в общем, так как он, очевидно, справедлив при $n = 1$.

Элементарные унитарные преобразования

43. В предыдущем параграфе мы использовали элементарные матрицы для того, чтобы преобразовать произвольный ненулевой вектор в ke_1 . Сейчас мы опишем первый из двух классов элементарных *унитарных* матриц, которые можно использовать для той же цели.

Рассмотрим матрицу R такую, что

$$\left. \begin{aligned} r_{ii} &= e^{i\alpha} \cos \theta, & r_{ij} &= e^{i\beta} \sin \theta, \\ r_{ji} &= e^{i\gamma} \sin \theta, & r_{jj} &= e^{i\delta} \cos \theta, \end{aligned} \right\} \quad (43.1)$$

$r_{pq} = 1$ ($p \neq i, j$), $r_{pq} = 0$ для остальных p и q , где $\theta, \alpha, \beta, \gamma, \delta$ вещественны. Эта матрица унитарна, если

$$\alpha - \gamma = \beta - \delta + \pi \pmod{2\pi}, \quad (43.2)$$

что можно проверить, составив $R^H R$.

Мы не нуждаемся в полной общности, которую дает соотношение (43.2), и будем считать

$$\gamma = \pi - \beta, \quad \delta = -\alpha, \quad (43.3)$$

откуда

$$\left. \begin{aligned} r_{ii} &= e^{i\alpha} \cos \theta, & r_{ij} &= e^{i\beta} \sin \theta, \\ r_{ji} &= -e^{-i\beta} \sin \theta, & r_{jj} &= e^{-i\alpha} \cos \theta. \end{aligned} \right\} \quad (43.4)$$

Такие матрицы мы будем называть *плоскими вращениями*. Ясно, что в силу (43.4) мы можем написать

$$r_{ii} = \bar{c}, \quad r_{ij} = \bar{s}, \quad r_{ji} = -s, \quad r_{jj} = c, \quad (43.5)$$

где $|c|^2 + |s|^2 = 1$.

Очевидно, R^H тоже является плоским вращением при α' и β' , удовлетворяющим соотношениям

$$\alpha' = -\alpha, \quad \beta' = \beta + \pi. \quad (43.6)$$

Поэтому обратная матрица к плоскому вращению тоже есть плоское вращение. При $\alpha = \beta = 0$ матрица вещественна и поэтому ортогональна, и связь с вращением на угол θ в плоскости (i, j) очевидна.

44. Если x — произвольный вектор, то для любых i и j мы можем выбрать плоское вращение такое, что компонента $(Rx)_i$ вещественна и не отрицательна, а компонента $(Rx)_j$ равна нулю. Ясно, что

$(Rx)_s = x_s$ для всех $s \neq i, j$. Мы имеем

$$(Rx)_j = -sx_i + cx_j. \quad (44.1)$$

Если $r^2 = |x_i|^2 + |x_j|^2$, то, предполагая, что r не нуль, можно взять

$$s = \frac{x_j}{r}, \quad c = \frac{x_i}{r}, \quad (44.2)$$

где r выбран положительным. Тогда ясно, что $(Rx)_j = 0$ и

$$(Rx)_i = \bar{c}x_i + \bar{s}x_j = \frac{|x_i|^2}{r} + \frac{|x_j|^2}{r} = r > 0. \quad (44.3)$$

Если $r = 0$, то можно взять $c = 1$ и $s = 0$. Если мы не нуждаемся в том, чтобы $(Rx)_i$ была вещественной, то мы можем опустить параметр α в уравнении (43.1). Заметим, что если x_i вещественна, то уравнение (44.2) дает вещественное значение для c в любом случае.

Вектор x может быть преобразован в ke_1 последовательным умножением слева на $(n-1)$ плоское вращение в плоскостях $(1, 2), (1, 3), \dots, (1, n)$. Вращение в $(1, i)$ -й плоскости выбирается так, чтобы свести i -й элемент к нулю. Как видно из доказательства, вращения можно выбирать так, что k вещественно и неотрицательно. Очевидно,

$$k = (|x_1|^2 + |x_2|^2 + \dots + |x_n|^2)^{1/2} = (x^H x)^{1/2}. \quad (44.4)$$

Отсюда следует, что если x и y таковы, что $x^H x = y^H y$, то существует произведение вращений S такое, что

$$Sx = y. \quad (44.5)$$

Действительно, как x , так и y могут быть преобразованы в ke_1 произведением матриц плоских вращений, а обратная матрица плоского вращения тоже задает плоское вращение.

Элементарные унитарные эрмитовы матрицы (матрицы отражения)

45. Второй класс элементарных унитарных матриц состоит из матриц P вида

$$P = I - 2ww^H, \quad w^H w = 1. \quad (45.1)$$

Эти матрицы унитарны, так как мы имеем

$$PP^H = (I - 2ww^H)(I - 2ww^H) = I - 4ww^H + 4w(w^H w)w^H = I. \quad (45.2)$$

Очевидно, они эрмитовы. Мы будем называть матрицы этого типа *элементарными эрмитовыми матрицами* *). Если w — вещественный вектор, то P ортогональна и симметрична.

Если y и x удовлетворяют соотношению

$$y = Px = x - 2w(w^H x), \quad (45.3)$$

то

$$y^H y = x^H P^H P x = x^H x \quad (45.4)$$

*) В русской литературе они названы матрицами отражения. Мы будем в дальнейшем придерживаться этого термина. (Прим. перев.)

и

$$x^H y = x^H P x, \quad (45.5)$$

и правая часть (45.5) вещественна, так как P эрмитова. Следовательно, мы не можем ожидать, что сможем найти такую P , для которой $y = Px$ при произвольных x и y . Если же выполнены условия $x^H x = y^H y$, и $x^H y$ вещественно, то, как мы сейчас покажем, мы можем найти такую P .

Если $x = y$, то любой вектор w такой, что $w^H x = 0$ даст подходящую P . С другой стороны, мы видим из (45.3), что каждый подходящий вектор w должен лежать на направлении $y - x$, и, следовательно, мы можем рассматривать в качестве вектора w только вектор вида

$$w = e^{i\alpha} (y - x) / [(y - x)^H (y - x)]^{1/2}. \quad (45.6)$$

Очевидно, что множитель $e^{i\alpha}$ не влияет на соответствующую P . То, что любой такой w подходит, можно легко доказать. Действительно,

$$(I - 2ww^H)x = x - \frac{2(y - x)(y - x)^H x}{(y - x)^H (y - x)} \quad (45.7)$$

и

$$\begin{aligned} 2(y - x)^H x &= 2(y^H x - x^H x) = y^H x + x^H y - x^H x - y^H y = \\ &= -(y - x)^H (y - x), \end{aligned} \quad (45.8)$$

так как в силу нашего предположения $x^H y$ вещественно и $x^H x = y^H y$. Следовательно,

$$(I - 2ww^H)x = x + (y - x) = y. \quad (45.9)$$

46. Вообще говоря, умножением слева на P невозможно превратить x в ke_1 с вещественным k . Если мы обозначим компоненты x через $r_i e^{i\theta_i}$, мы можем превратить x в y вида

$$y = \pm (x^H x)^{1/2} e^{i\theta_1} e_1, \quad (46.1)$$

так как тогда очевидно, что $x^H y$ вещественно. Для соответствующей P мы можем взять

$$w = (x - y) / [(x - y)^H (x - y)]^{1/2}. \quad (46.2)$$

Если $r_1 = 0$, то θ_1 может быть произвольным, и если мы возьмем $\theta_1 = 0$, то соответствующая P переведет x в ke_1 , где k вещественно. Мы часто будем заинтересованы также в таких P , для которых соответствующий вектор w имеет нулевые компоненты на первых r позициях, при r от 0 до $n - 1$. Умножение слева вектора x на такую P оставляет первые r компонент x неизменными.

Если x — вещественный, то y и w в (46.1) и (46.2) вещественны и, следовательно, P вещественна. Заметим, что с помощью унитарной матрицы вида $e^{i\theta} P$ мы можем перевести произвольный вектор в вектор вида ke_1 с вещественным k .

Приведение к треугольному виду элементарными унитарными преобразованиями

47. Сейчас докажем два результата, сформулированные ранее, о приведении матрицы к канонической форме ортогональными (унитарными) преобразованиями (§§ 23, 25). Сначала покажем, что любую матрицу можно

привести унитарным преобразованием к треугольной форме. Доказательство проведем по индукции.

Пусть это верно для матриц $(n - 1)$ -го порядка. Пусть A — матрица n -го порядка, а λ и x — ее собственное значение и собственный вектор, так что

$$Ax = \lambda x. \quad (47.1)$$

Мы знаем из §§ 44 и 46, что существует унитарная матрица P такая, что

$$Px = ke_1, \quad (47.2)$$

где P может быть либо произведением $(n - 1)$ матриц плоских вращений, либо одной матрицей отражения. Из (47.1) следует, что

$$PAx = P\lambda x, \quad PA(P^H P)x = \lambda Px, \quad (47.3)$$

$(PAR^H)ke_1 = \lambda ke_1$, $(PAR^H)e_1 = \lambda e_1$, так как $k \neq 0$. Матрица PAR^H должна поэтому иметь вид

$$A^{(1)} = \begin{bmatrix} \lambda & b^T \\ 0 & A_{n-1} \end{bmatrix}, \quad (47.4)$$

где A_{n-1} матрица порядка $(n - 1)$. Теперь, по предположению, существует унитарная матрица R такая, что

$$RA_{n-1}R^H = T_{n-1}, \quad (47.5)$$

где T_{n-1} — треугольная. Следовательно,

$$\begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} A^{(1)} \begin{bmatrix} 1 & 0 \\ 0 & R^H \end{bmatrix} = \begin{bmatrix} \lambda & b^T R^H \\ 0 & T_{n-1} \end{bmatrix}, \quad (47.6)$$

и матрица в правой части (47.6) треугольная. Так как $\begin{bmatrix} 1 & 0 \\ 0 & R \end{bmatrix} P$ — унитарная матрица, результат установлен. Заметим, что если A вещественна и ее собственные значения вещественны, то ее собственные векторы также вещественны, и, следовательно, матрица P ортогональная, и все преобразование можно осуществить, используя ортогональные матрицы.

Если A — эрмитова, треугольная матрица диагональна. Действительно, существует унитарная матрица R такая, что

$$RAR^H = T, \quad (47.7)$$

где T — треугольная. Матрица слева эрмитова, поэтому T — эрмитова, и, следовательно, диагональная.

Нормальные матрицы

48. Мы доказали, что для любой эрмитовой матрицы существует унитарная матрица R такая, что

$$RAR^H = D, \quad (48.1)$$

где D — диагональная и вещественная. Сейчас мы рассмотрим расширение этого класса матриц, которое получается, если позволить D быть комплексной. Другими словами, мы рассмотрим матрицы A , которые могут быть представлены в виде $R^H D R$, где R — унитарная и D — диагональная. Если

$$A = R^H D R, \quad (48.2)$$

то

$$AA^H = R^H D R R^H D^H R = R^H (DD^H) R, \quad (48.3)$$

$$A^H A = R^H D^H R R^H D R = R^H (D^H D) R. \quad (48.4)$$

Следовательно, $AA^H = A^H A$, так как DD^H , очевидно, равно DD^H .

Покажем обратное, т. е. покажем, что если

$$AA^H = A^H A, \quad (48.5)$$

то A можно факторизовать, как в (48.2). Действительно, любая A может быть представлена в виде $R^H T R$, где R — унитарная, а T — верхняя треугольная. Следовательно, мы имеем

$$R^H T R R^H T^H R = R^H T^H R R^H T R,$$

что дает

$$R^H T T^H R = R^H T^H T R. \quad (48.6)$$

Умножая слева на R и справа на R^H , получаем

$$T T^H = T^H T. \quad (48.7)$$

Приравнивая диагональные элементы, получаем, что все недиагональные элементы T — нули, так что T — диагональная. Следовательно, класс матриц, которые можно факторизовать в виде $R^H D R$, совпадает с классом матриц, для которых $AA^H = A^H A$. Такие матрицы называются *нормальными* матрицами. Из соотношения (48.5) очевидно, что в класс нормальных матриц входят матрицы, для которых

- (i) $A^H = A$ (эрмитовы матрицы),
- (ii) $A^H = -A$ (косоэрмитовы матрицы),
- (iii) $A^H = A^{-1}$ (унитарные матрицы).

Из практики мы заметили, что задача об отыскании собственных значений косоэрмитовых и унитарных матриц возникает сравнительно редко.

Коммутирующие матрицы

49. Нормальные матрицы обладают тем свойством, что A и A^H коммутируют. Из соотношения

$$A = R^H D R \quad (49.1)$$

имеем

$$A R^H = R^H D, \quad (49.2)$$

и аналогично

$$A^H R^H = R^H D^H. \quad (49.3)$$

Матрицы A и A^H поэтому имеют общую полную систему собственных векторов, например ту, что образуют столбцы матрицы R^H .

Мы сейчас покажем, что если две матрицы A и B имеют общую полную систему собственных векторов, то $AB = BA$. Действительно, эта система собственных векторов образует столбцы матрицы H , и мы имеем

$$A = H D_1 H^{-1}, \quad B = H D_2 H^{-1}. \quad (49.4)$$

Следовательно,

$$AB = H D_1 H^{-1} H D_2 H^{-1} = H D_1 D_2 H^{-1}, \quad (49.5)$$

и

$$BA = HD_2H^{-1}HD_1H^{-1} = HD_2D_1H^{-1}, \quad (49.6)$$

так что $AB = BA$, потому что диагональные матрицы коммутируют.

50. Обратно, если A и B коммутируют и имеют линейные элементарные делители, то они имеют общую систему собственных векторов. Действительно, предположим, что $h_1, h_2, \dots, h_n$ — собственные векторы A и что они образуют матрицу H . Тогда

$$H^{-1}AH = \text{diag}(\lambda_i), \quad (50.1)$$

где λ_i могут быть или не быть различными. Структура доказательства будет ясна, если мы рассмотрим матрицу восьмого порядка, для которой $\lambda_1 = \lambda_2 = \lambda_3, \lambda_4 = \lambda_5$, а λ_6, λ_7 и λ_8 изолированные. Уравнение (50.1) можно записать в виде

$$H^{-1}AH = \begin{bmatrix} \lambda_1 I_3 & & & \\ & \lambda_4 I_2 & & \\ & & \lambda_6 & \\ & & & \lambda_7 \\ & & & & \lambda_8 \end{bmatrix}. \quad (50.2)$$

Теперь, так как $Ah_i = \lambda_i h_i$, имеем

$$BAh_i = \lambda_i Bh_i, \quad (50.3)$$

что дает

$$A(Bh_i) = \lambda_i(Bh_i). \quad (50.4)$$

Это показывает, что если h_i — собственный вектор A , соответствующий λ_i , то Bh_i лежит в подпространстве, натянутом на собственные векторы, соответствующие λ_i . Следовательно,

$$\left. \begin{aligned} Bh_1 &= p_{11}h_1 + p_{21}h_2 + p_{31}h_3, \\ Bh_2 &= p_{12}h_1 + p_{22}h_2 + p_{32}h_3, \\ Bh_3 &= p_{13}h_1 + p_{23}h_2 + p_{33}h_3, \\ Bh_4 &= q_{11}h_4 + q_{21}h_5, \\ Bh_5 &= q_{12}h_4 + q_{22}h_5, \end{aligned} \right\} \begin{aligned} Bh_6 &= \mu_6 h_6, \\ Bh_7 &= \mu_7 h_7, \\ Bh_8 &= \mu_8 h_8. \end{aligned} \quad (50.5)$$

Эти уравнения эквивалентны одному матричному уравнению

$$H^{-1}BH = \begin{bmatrix} P & & & \\ & Q & & \\ & & \mu_6 & \\ & & & \mu_7 \\ & & & & \mu_8 \end{bmatrix} \quad (50.6)$$

Так как матрица B имеет линейные элементарные делители, это же самое должно быть верно и для матрицы в правой части уравнения (50.6), и, следовательно, для P и Q . Поэтому существуют матрицы K и L третьего и второго порядков такие, что

$$K^{-1}PK = \begin{bmatrix} \mu_1 & & \\ & \mu_2 & \\ & & \mu_3 \end{bmatrix}, \quad L^{-1}QL = \begin{bmatrix} \mu_4 & \\ & \mu_5 \end{bmatrix}. \quad (50.7)$$

Если мы положим

$$G = \begin{bmatrix} K & & \\ & L & \\ & & I_3 \end{bmatrix}, \quad (50.8)$$

то

$$G^{-1}H^{-1}BHG = \text{diag}(\mu_i). \quad (50.9)$$

С другой стороны, из (50.2) мы видим, что

$$G^{-1}H^{-1}AHG = \text{diag}(\lambda_i). \quad (50.10)$$

Следовательно, B и A имеют общую систему собственных векторов, например векторов, являющихся столбцами HG .

Собственные значения AB

51. Заметим, что собственные значения AB и BA одинаковы для всех квадратных A и B . Доказательство следующее. Мы имеем

$$\begin{bmatrix} I & 0 \\ -B & \mu I \end{bmatrix} \begin{bmatrix} \mu I & A \\ B & \mu I \end{bmatrix} = \begin{bmatrix} \mu I & A \\ 0 & \mu^2 I - BA \end{bmatrix} \quad (51.1)$$

и

$$\begin{bmatrix} \mu I & -A \\ 0 & I \end{bmatrix} \begin{bmatrix} \mu I & A \\ B & \mu I \end{bmatrix} = \begin{bmatrix} \mu^2 I - AB & 0 \\ B & \mu I \end{bmatrix}. \quad (51.2)$$

Сосчитав определители обеих частей (51.1) и (51.2) и полагая

$$\begin{bmatrix} \mu I & A \\ B & \mu I \end{bmatrix} = X, \quad (51.3)$$

получим

$$\mu^n \det(X) = \mu^n \det(\mu^2 I - BA) \quad (51.4)$$

и

$$\mu^n \det(X) = \mu^n \det(\mu^2 I - AB). \quad (51.5)$$

Уравнения (51.4) и (51.5) суть тождества по μ^2 , и обозначив $\mu^2 = \lambda$, получим

$$\det(\lambda I - BA) = \det(\lambda I - AB), \quad (51.6)$$

что показывает, что собственные значения AB и BA одинаковы.

То же самое доказательство показывает, что если A есть $(m \times n)$ -матрица, а B есть $(n \times m)$ -матрица, то AB и BA имеют одинаковые собственные значения, за исключением того, что произведение большего порядка имеет сверх того $|n - m|$ нулевых собственных значений.

Векторные и матричные нормы

52. Как векторам, так и матрицам оказывается удобно сопоставлять некоторое число, характеризующее в некотором смысле их величину и играющее ту же роль, что и модуль для комплексных чисел. Для этой цели будем использовать определенные функции элементов вектора или матрицы, которые называются *нормами*. *Норма вектора* будет обозна-

чаться $\|x\|$, и все нормы, которыми мы пользуемся, удовлетворяют соотношениям

$$\left. \begin{aligned} \text{(i)} \quad & \|x\| > 0 \text{ при } x \neq 0 \text{ и } \|0\| = 0, \\ \text{(ii)} \quad & \|kx\| = |k| \|x\| \text{ для любого комплексного скаляра } k, \\ \text{(iii)} \quad & \|x + y\| \leq \|x\| + \|y\|. \end{aligned} \right\} \quad (52.1)$$

Из (iii) имеем

$$\|x - y\| \geq |\|x\| - \|y\||.$$

Мы будем употреблять только три векторные нормы, именно

$$\|x\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p} \quad (p = 1, 2, \infty), \quad (52.2)$$

где $\|x\|_\infty$ интерпретируется как $\max |x_i|$.

Норма $\|x\|_2$ есть евклидова длина вектора x в обычном смысле. Мы также имеем

$$x^H x = \|x\|_2^2 \quad (52.3)$$

и

$$|x^H y| \leq \sum |x_i| |y_i| \leq \|x\|_2 \|y\|_2 \quad (52.4)$$

по теореме Коши. Поэтому можно записать

$$x^H y = \|x\|_2 \|y\|_2 e^{i\theta} \cos \theta \quad \left(0 \leq \theta \leq \frac{\pi}{2}\right). \quad (52.5)$$

Если x и y вещественны, то θ — угол между двумя векторами; в комплексном случае мы можем рассматривать θ , определенный в (52.5), как обобщение угла между двумя векторами.

Аналогично нормы матрицы A будем обозначать $\|A\|$, и все употребляемые нами нормы удовлетворяют соотношениям

$$\left. \begin{aligned} \text{(i)} \quad & \|A\| > 0 \text{ при } A \neq 0 \text{ и } \|0\| = 0, \\ \text{(ii)} \quad & \|kA\| = |k| \|A\| \text{ для любого комплексного} \\ & \text{скаляра } k, \\ \text{(iii)} \quad & \|A + B\| \leq \|A\| + \|B\|, \\ \text{(iv)} \quad & \|AB\| \leq \|A\| \|B\|. \end{aligned} \right\} \quad (52.6)$$

Подчиненные матричные нормы

53. Каждой матрице A можно сопоставить неотрицательную величину $\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$, соответствующую любой векторной норме. Из соотношения (ii) (52.1) видим, что это эквивалентно $\sup_{\|x\|=1} \|Ax\|$. Эта величина, очевидно, является функцией матрицы A , и легко проверить, что она удовлетворяет всем условиям, которым должна удовлетворять матричная норма. Она называется матричной нормой, *подчиненной* векторной норме. Так как

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \quad (53.1)$$

мы имеем

$$\|Ax\| \leq \|A\| \|x\| \quad (53.2)$$

для всех отличных от нуля x , и соотношение, очевидно, справедливо, когда $x = 0$. Матричные и векторные нормы, для которых справедливо (53.2) для всех A и x , называются *согласованными*. Векторная норма и подчиненная ей матричная норма поэтому всегда согласованы.

Для всех подчиненных норм имеем по определению $\|I\| = 1$.

Далее, всегда существует ненулевой вектор такой, что

$$\|Ax\| = \|A\| \|x\|.$$

Это следует из замкнутости области $\|x\| = 1$ и того, что в силу (52.1) векторная норма является непрерывной функцией компонент своего аргумента. Поэтому вместо $\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ можем писать $\max \frac{\|Ax\|}{\|x\|}$.

Матричная норма, подчиненная $\|x\|_p$, обозначается $\|A\|_p$. Эти нормы удовлетворяют соотношениям

$$\|A\|_1 = \max_i \sum_j |a_{ij}|, \quad (53.3)$$

$$\|A\|_\infty = \max_j \sum_i |a_{ij}|, \quad (53.4)$$

$$\|A\|_2 = (\text{наибольшее собственное значение } A^H A)^{1/2}. \quad (53.5)$$

Первые два результата тривиальны, и ясно, что $\|A\|_1 = \|A^H\|_\infty$. Третий можно проверить следующим образом.

Матрица $A^H A$ эрмитова, и ее собственные значения не отрицательны, так как $x^H A^H A x = (Ax)^H (Ax) \geq 0$. Обозначим ее собственные значения через $\sigma_1^2, \sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_n^2 \geq 0$. По определению

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}, \quad (53.6)$$

$$\|A\|_2^2 = \max_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{x^H A^H A x}{x^H x}. \quad (53.7)$$

Далее, если $u_1, u_2, \dots, u_n$ — ортонормированная система собственных векторов $A^H A$, то можно написать

$$x = \sum \alpha_i u_i. \quad (53.8)$$

Следовательно,

$$\frac{x^H A^H A x}{x^H x} = \frac{\sum |\alpha_i|^2 \sigma_i^2}{\sum |\alpha_i|^2} \leq \sigma_1^2. \quad (53.9)$$

Величина σ_1^2 достигается при $x = u_1$. Неотрицательные величины σ_i называются *сингулярными (главными) значениями* A . Норма $\|A\|_2$ часто называется *спектральной нормой*.

Евклидова и спектральная нормы

54. Существует вторая важная норма, согласованная с векторной нормой $\|x\|_2$. Это так называемая *евклидова или шуровская* норма, которую мы обозначим $\|A\|_E$. По определению

$$\|A\|_E = \left(\sum_{ij} |a_{ij}|^2 \right)^{1/2}. \quad (54.1)$$

Очевидно, что она не может быть подчинена никакой векторной норме, так как

$$\|I\|_E = n^{1/2}. \quad (54.2)$$

Евклидова норма весьма часто используется для практических целей, так как ее легко можно сосчитать. Кроме того, имеем по определению

$$\| \|A\| \|_E = \|A\|_1 \quad (54.3)$$

Нормы $\|\cdot\|_1$ и $\|\cdot\|_\infty$ удовлетворяют таким же соотношениям, но, вообще говоря,

$$\| \|A\| \|_2 \neq \|A\|_2. \quad (54.4)$$

В самом деле, так как $\|A\|_E^2$ равна следу $A^H A$, который равен $\sum \sigma_i^2$, имеем

$$\|A\|_2 \leq \|A\|_E \leq n^{1/2} \|A\|_2 \quad (54.5)$$

где обе границы могут быть достигнуты. Следовательно, мы имеем

$$\| \|A\| \|_2 \leq \| \|A\| \|_E = \|A\|_E \leq n^{1/2} \|A\|_2. \quad (54.6)$$

Неудобством евклидовой нормы является то, что

$$\|\text{diag}(\lambda_i)\|_E = (\sum |\lambda_i|^2)^{1/2}, \quad (54.7)$$

в то время как

$$\|\text{diag}(\lambda_i)\|_p = \max |\lambda_i| \quad (p = 1, 2, \infty). \quad (54.8)$$

Как евклидова матричная норма, так и вторая векторная и вторая матричная нормы имеют важные инвариантные свойства при унитарных преобразованиях. Для произвольных x и A и унитарной R имеем

$$\|Rx\|_2 = \|x\|_2, \quad (54.9)$$

так как

$$\|Rx\|_2^2 = (Rx)^H Rx = x^H R^H Rx = x^H x = \|x\|_2^2. \quad (54.10)$$

Аналогично из определения следует, что

$$\|RA\|_2 = \|A\|_2 \quad \text{и} \quad \|RAR^H\|_2 = \|A\|_2. \quad (54.11)$$

Наконец,

$$\|RA\|_E = \|A\|_E, \quad (54.12)$$

ибо евклидова длина каждого столбца RA равна длине соответствующего столбца A .

Предположим, что $A = R^H \text{diag}(\lambda_i) R$ — нормальная матрица. Тогда

$$\begin{aligned} \|AB\|_E &= \|R^H \text{diag}(\lambda_i)^T RB\|_E = \|\text{diag}(\lambda_i, RB)\|_E \leq \\ &\leq \max |\lambda_i| \|RB\|_E = \max |\lambda_i| \|B\|_E. \end{aligned} \quad (54.13)$$

В дальнейшем под нормой будет подразумеваться одна из специальных норм, которые мы сейчас определили, и мы не будем сколько-нибудь обобщать это понятие. Однако даже при таком узком определении оказывается удобным иметь единообразную трактовку норм, и удивительно, как много выигрывается от легкости их употребления.

Нормы и пределы

55. Установим несколько простых критериев сходимости последовательностей матриц. Там, где результат немедленно следует из определения, доказательства опущены.

(А) $\lim_{r \rightarrow \infty} A^r = A$ тогда и только тогда, когда $\|A - A^{(r)}\| \rightarrow 0$. Если $A^{(r)} \rightarrow A$, то $\|A^r\| \rightarrow \|A\|$.

(В) $\lim_{r \rightarrow \infty} A^r = 0$, если $\|A\| < 1$.

Действительно,

$$\|A^r\| = \|AA^{r-1}\| \leq \|A\| \|A^{r-1}\| \leq \|A\|^2 \|A^{r-2}\| \leq \dots \leq \|A\|^r.$$

(С) Если λ — собственное значение A , то $|\lambda| \leq \|A\|$.

Действительно, мы имеем $Ax = \lambda x$ для некоторого отличного от нуля x . Следовательно, для любой согласованной пары матричной и векторной норм

$$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \|x\|,$$

и

$$|\lambda| \leq \|A\|. \quad (55.1)$$

Отсюда имеем

$$\|A\|_2^2 = \text{наибольшее собственное значение } A^H A \leq \|A^H A\|_1 \leq \|A^H\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1. \quad (55.2)$$

(D) $\lim_{r \rightarrow \infty} A^r = 0$ тогда и только тогда, когда $|\lambda_i| < 1$ для всех собственных значений.

Это следует из рассмотрения жордановой канонической формы. Для типичной жордановой подматрицы имеем

$$\begin{bmatrix} a & & \\ & 1 & a \\ & & 1 & a \end{bmatrix}^r = a^{r-2} \begin{bmatrix} a^2 & & \\ ra & a^2 & \\ \frac{1}{2} r(r-1) & ra & a^2 \end{bmatrix}, \quad (55.3)$$

и ясно, что эта матрица стремится к нулю тогда и только тогда, когда $|a| < 1$. Обычно доказывают результат (С) из (D), хотя, если мы интересуемся лишь согласованными парами норм, доказательство, данное нами, более естественно. Величина

$$\rho(A) = \max |\lambda_i| \quad (55.4)$$

называется *спектральным радиусом* A . Мы видели, что

$$\rho(A) \leq \|A\| \quad (55.5)$$

(Е) Ряд $I + A + A^2 + \dots$ сходится тогда и только тогда, когда $A^r \rightarrow 0$.

Достаточность этого условия следует из (D). Действительно, если оно выполнено, то все собственные значения A лежат в единичном круге, и, следовательно, $(I - A)$ неособенная. Соотношение

$$(I - A)(I + A + A^2 + \dots + A^r) = I - A^{r+1} \quad (55.6)$$

выполняется тождественно, так что

$$I + A + A^2 + \dots + A^r = (I - A)^{-1} - (I - A)^{-1} A^{r+1}. \quad (55.7)$$

Второй член в правой части стремится к нулю, и поэтому

$$I + A + A^2 + \dots \rightarrow (I - A)^{-1}.$$

Необходимость условия очевидна.

(F) Для того чтобы ряд $I + A + A^2 + \dots$ сходиллся, достаточно, чтобы какая-либо из норм A была меньше единицы.

Действительно, если это так, то мы имеем из (C)

$$|\lambda_i| < 1$$

для всех собственных значений. Следовательно, в силу (D) $A^r \rightarrow 0$, и поэтому в силу (E) ряд сходится.

Заметим, что существование такой нормы, что $\|A\| > 1$, может не препятствовать сходимости ряда. Например, если

$$A = \begin{bmatrix} 0,9 & 0 \\ 0,3 & 0,8 \end{bmatrix}, \quad (55.8)$$

то $\|A\|_\infty = 1,1$, $\|A\|_1 = 1,2$ и $\|A\|_E = (1,54)^{1/2}$, а $A^r \rightarrow 0$, так как ее собственные значения равны 0,9 и 0,8.

При попытках установить сходимость ряда на практике мы свободны в выборе наименьшей нормы.

(G) Для того чтобы ряд из (E) сходиллся, достаточно, чтобы какая-либо норма любого преобразования подобия A была меньше единицы. Действительно, если $A = H B H^{-1}$, то

$$I + A + A^2 + \dots + A^r = H(I + B + \dots + B^r)H^{-1} \quad (55.9)$$

Ряд справа сходится, если $\|B\| < 1$.

Мы можем использовать преобразование подобия для получения матрицы, у которой одна из наших стандартных норм меньше, чем у исходной матрицы. Для матрицы, данной в (F), мы можем взять

$$B = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} A \begin{bmatrix} 1/3 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0,9 & 0 \\ 0,1 & 0,8 \end{bmatrix}, \quad (55.10)$$

и тогда $\|B\|_\infty = 0,9$.

Рассмотрим более убедительный пример

$$A = \begin{bmatrix} 0,1 & 0 & 1,0 \\ 0,2 & 0,3 & 0 \\ 0,2 & 0,1 & 0,4 \end{bmatrix}. \quad (55.11)$$

Имеем $\|A\|_\infty = 1,1$, $\|A\|_1 = 1,4$, $\|A\|_E = (1,35)^{1/2}$. Определим

$$B = \begin{bmatrix} 0,5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} A \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0,1 & 0 & 0,5 \\ 0,4 & 0,3 & 0 \\ 0,4 & 0,1 & 0,4 \end{bmatrix} \quad (55.12)$$

Мы видим, что $\|B\|_\infty = 0,9$, так что ряд из A сходится. Преобразование подобия с диагональной матрицей часто играет большую роль в анализе ошибок.

Некоторые оценки без использования бесконечных матричных рядов

56. Мы можем часто, используя элементарные свойства норм, избегать применения бесконечных рядов матриц. Например, мы можем получить оценку для $\|(I + A)^{-1}\|$, если $\|A\| < 1$, следующим образом.

Так как $\|A\| < 1$, все собственные значения A меньше, чем единица. Поэтому $I + A$ неособенная и $(I + A)^{-1}$ существует. Мы имеем

$$I = (I + A)^{-1}(I + A) = (I + A)^{-1} + (I + A)^{-1}A. \quad (56.1)$$

Следовательно,

$$\begin{aligned} \|I\| &\geq \|(I + A)^{-1}\| - \|(I + A)^{-1}A\| \geq \\ &\geq \|(I + A)^{-1}\| - \|(I + A)^{-1}\| \|A\|, \end{aligned} \quad (56.2)$$

$$\|(I + A)^{-1}\| \leq \frac{\|I\|}{(1 - \|A\|)} = (1 - \|A\|)^{-1}, \quad \text{если } \|I\| = 1. \quad (56.3)$$

С другой стороны, можем написать

$$(I + A)^{-1} = 1 - A + A^2 - \dots, \quad (56.4)$$

причем сходимость гарантирована, если $\|A\| < 1$. Из (56.4) заключаем

$$\|(I + A)^{-1}\| \leq \|I\| + \|A\| + \|A\|^2 + \dots = (1 - \|A\|)^{-1}, \quad \text{если } \|I\| = 1. \quad (56.5)$$

Дополнительные замечания

В этой главе я старался дать обзор основного материала, который понадобится в остальной части книги. Серьезная исследовательская работа в этом направлении требует, помимо чтения этой главы, ознакомления с дополнительной литературой. Очень полное изложение классической теории содержится в «Теории матриц» Гантамахера (1967). В библиографию я включил ряд стандартных книг по этому вопросу. «The theory of matrices in numerical analysis» Хаусхолдера (1964) заслуживает специального упоминания, так как в ней в дополнение к основному материалу этой главы фактически имеется весь теоретический материал, обсуждаемый в остальных главах.

ТЕОРИЯ ВОЗМУЩЕНИЙ

Введение

1. В первой главе мы имели дело с математическим обоснованием алгебраической проблемы собственных значений. Остальная часть книги посвящена рассмотрению практических задач, возникающих при вычислении собственных значений и собственных векторов на цифровых вычислительных машинах и в определении точности вычисленных значений. Главное внимание будет обращено на оценку влияния различных ошибок, которые присущи формулировке задачи и способам ее решения. Здесь возникают три главных проблемы.

(i) Элементы данной матрицы могут определяться прямо из физических наблюдений, и потому они могут иметь ошибки, свойственные всем наблюдениям. В этом случае матрица A , которую мы имеем, является приближением матрицы, соответствующей точным измерениям. Если мы можем утверждать, что ошибка в любом элементе A ограничена числом δ , то можем сказать, что правильная матрица есть $(A + E)$, где E — некоторая матрица, для которой

$$|e_{ij}| \leq \delta. \quad (1.1)$$

Полное решение поставленной задачи состоит не только в определении собственных значений A , но также и в оценке возможных вариаций собственных значений всех матриц класса $(A + E)$, где E удовлетворяет условию (1.1). Мы сразу приходим к задаче о нахождении возмущений собственных значений матрицы, соответствующих возмущению ее элементов.

(ii) Элементы матрицы могут быть заданы точно математическими формулами, но работать с такой точной матрицей на цифровых машинах все же невозможно. Если элементы матрицы иррациональны, это почти всегда так. Но даже если они рациональны, их точное представление может потребовать большего числа цифр, чем имеется в машинной шкале. Матрица A , например, может быть произведением некоторых матриц, элементы которых имеют полное число знаков, нормально используемое в вычислительной машине. Очевидно, что мы имеем дело с почти той же проблемой, что и в (i). Данная матрица будет $(A + E)$, где E — матрица ошибки.

(iii) Даже если мы можем рассматривать матрицу, заданную в вычислительной машине, как точную, это все же неверно, вообще говоря, для вычисленного решения. Наиболее часто при вычислении решения считают последовательность преобразований подобия $A_1, A_2, A_3, \dots$ исходной матрицы A , и ошибки округления вносятся при проведении каждого из этих преобразований. На первый взгляд ошибки, возникающие из этих преобразований, имеют другую природу, чем ошибки в (i) и (ii). Однако, вообще говоря, это неверно. Часто мы сможем показать, что вычисленные матрицы A_i точно подобны матрицам $(A + E_i)$, где E_i имеют малые элементы, являющиеся функциями ошибок округления.

а также найти точные границы для E_i . Собственные значения A_i будут поэтому точно равны собственным значениям $(A + E_i)$, и снова мы приходим к рассмотрению возмущений собственных значений и собственных векторов матрицы, соответствующих возмущению ее элементов.

Теорема Островского о непрерывности собственных значений

2. Верхние границы возмущений собственных значений были даны Островским (1957) в следующем виде.

Пусть A и B — матрицы, элементы которых удовлетворяют соотношениям

$$|a_{ij}| < 1, \quad |b_{ij}| < 1. \quad (2.1)$$

Тогда, если λ' — собственное значение $(A + \varepsilon B)$, то существует собственное значение λ матрицы A такое, что

$$|\lambda' - \lambda| < (n+2)(n^2\varepsilon)^{1/n} \quad (2.2)$$

Кроме того, собственные значения λ и λ' матриц A и $(A + \varepsilon B)$ могут быть приведены во взаимно однозначное соответствие так, что

$$|\lambda' - \lambda| < 2(n+1)^2 (n^2\varepsilon)^{1/n}. \quad (2.3)$$

Хотя эти результаты имеют большую теоретическую важность, они имеют малую практическую ценность. Предположим, что мы имеем матрицу A 20-го порядка с иррациональными элементами, удовлетворяющими соотношениям

$$|a_{ij}| < 1. \quad (2.4)$$

Рассмотрим матрицу $(A + E)$, полученную округлением элементов A до десяти десятичных знаков, так что

$$|e_{ij}| \leq \frac{1}{2} \cdot 10^{-10}, \quad (2.5)$$

где верхняя граница достижима. Из второго результата Островского

$$|\lambda' - \lambda| < 2 \times 21^2 \cdot \left(400 \times \frac{1}{2} \cdot 10^{-10} \right)^{1/20} \quad (2.6)$$

Эта граница значительно хуже, чем граница, которую можно получить из простой теории норм. Действительно,

$$\left. \begin{aligned} |\lambda| &\leq \|A\|_\infty < 20, \\ |\lambda'| &\leq \|A + E\|_\infty < 20 + 10 \times 10^{-10}, \end{aligned} \right\} \quad (2.7)$$

и, следовательно,

$$|\lambda' - \lambda| \leq |\lambda| + |\lambda'| < 40 + 10^{-9}. \quad (2.8)$$

Результаты Островского улучшают результаты, получаемые с помощью простой теории норм, только тогда, когда

$$2(n+1)^2 (n^2\varepsilon)^{1/n} < 2n,$$

что дает

$$\varepsilon < \frac{n^{n-2}}{(n+1)^{2n}} \approx \frac{1}{e^2 n^{n+2}}. \quad (2.9)$$

Для $n = 20$ это означает, что ε должно быть меньше 10^{-27} . Для получения *полезных* границ из результатов Островского число ε должно быть значительно меньше даже этого.

Очевидно, что множитель $\varepsilon^{1/n}$ является главной слабостью результата Островского. Однако, как показывают примеры Форсайта, его присутствие неизбежно в любом общем результате. Рассмотрим, например, $C_n(a)$ — жорданову подматрицу порядка n , у которой n собственных значений равны a . Если мы заменим нулевой элемент в позиции $(1, n)$ на ε , то характеристическое уравнение будет

$$(a - \lambda)^n + (-1)^{n-1} \varepsilon = 0; \quad (2.10)$$

что дает

$$\lambda = a + \omega^r \varepsilon^{1/n} \quad (r = 0, 1, \dots, n-1) \quad (2.11)$$

для n собственных значений, где ω — любой примитивный корень n -й степени из единицы.

Алгебраические функции

3. В дальнейшем нам понадобятся два результата из теории алгебраических функций (см., например, Гурса, 1933). Мы их сформулируем без доказательства.

Рассмотрим

$$f(x, y) = y^n + p_{n-1}(x) y^{n-1} + p_{n-2}(x) y^{n-2} + \dots + p_1(x) y + p_0(x), \quad (3.1)$$

где $p_i(x)$ — полиномы от x . При любом значении x уравнение $f(x, y) = 0$ имеет n корней $y_1(x), y_2(x), \dots, y_n(x)$, причем каждый корень взят столько раз, какова его кратность. Корни $f(0, y) = 0$ обозначим через $y_1(0), y_2(0), \dots, y_n(0)$. Справедливы две теоремы.

Т е о р е м а 1. *Предположим, что $y_i(0)$ — простой корень $f(0, y) = 0$. Тогда найдется положительное число δ_i такое, что существует простой корень $y_i(x)$ уравнения $f(x, y) = 0$ вида*

$$y_i(x) = y_i(0) + p_{i1}x + p_{i2}x^2 + \dots, \quad (3.2)$$

причем ряд в правой части (3.2) сходится при $|x| < \delta_i$.

З а м е ч а н и я. (i) Мы требуем, чтобы только корень $y_i(0)$, который мы рассматриваем, был простым; о природе остальных корней уравнения $f(0, y) = 0$ не делается никаких предположений.

(ii) Ряд в правой части (3.2) может обрываться.

(iii) $y_i(x) \rightarrow y_i(0)$ при $x \rightarrow 0$.

Т е о р е м а 2. *Если $y_1(0) = y_2(0) = \dots = y_m(0)$ — корень m -й кратности уравнения $f(0, y) = 0$, то найдется положительное число δ такое, что существуют точно m нулей $f(x, y) = 0$, которые при $|x| < \delta$ имеют следующие свойства.*

Эти m корней распадаются в r групп по $m_1, m_2, \dots, m_r$ корней соответственно, где

$$\sum m_i = m, \quad (3.3)$$

и корни в группе m_i определяются как m_i значений рядов

$$y_1(0) + p_{i1}z + p_{i2}z^2 + \dots, \quad (3.4)$$

соответствующих m_i различным значениям z вида

$$z = x^{m_i^{-1}}. \quad (3.5)$$

- З а м е ч а н и я. (i) Снова некоторые из рядов (3.4) могут обрываться.
(ii) Может быть, что $r = 1$, тогда $m_1 = m$ и все m корней даются одним и тем же дробным степенным рядом.
(iii) При некотором i может быть $m_i = 1$, так что в этом случае соответствующий степенной ряд не имеет дробных степеней.
(iv) Все m корней стремятся к $y_1(0)$ при x , стремящемся к нулю.

Численные примеры

4. (i) $f(x, y) = y^2(1+x) - 3y(1+x^2) + (2+x)$.

Имеем

$$f(0, y) = y^2 - 3y + 2.$$

Следовательно, например, $y_1(0) = 1$ — простой корень. Соответствующий корень $y_1(x)$ есть

$$y_1(x) = \frac{3(1+x^2) - (1 - 12x + 14x^2 + 9x^4)^{1/2}}{2(1+x)},$$

и его, очевидно, можно разложить в сходящийся степенной ряд по x , если $12|x| + 14|x|^2 + 9|x|^4 < 1$.

(ii) $f(x, y) = y^3 - y^2 - x(1+x)^2$.

Имеем $f(0, y) = y^3 - y^2$, так что $y_1(0) = 1$ — простой корень, а $y_2(0) = 0$ — двукратный. Очевидно,

$$y_1(x) = 1 + x,$$

так что сходящийся ряд по x обрывается. Заметим, что присутствие кратного корня не ведет к дробным степеням в $y_1(x)$.

Корни, соответствующие $y_2(0) = 0$, являются решениями уравнения

$$y^2 + yx + x(1+x) = 0$$

и поэтому даются двумя значениями функции

$$y = \frac{1}{2}[-x + (-4x - 3x^2)^{1/2}] = \frac{1}{2}\left[-x + 2ix^{1/2}\left(1 + \frac{3}{4}x\right)^{1/2}\right].$$

Степенные ряды поэтому содержат лишь нечетные степени $x^{1/2}$; они сходятся при $|x| < 4/3$.

(iii) $f(x, y) = y^4 - y^2x(1+x)^2 + x^3(1+x)^2$.

Уравнение $f(0, y) = 0$ имеет корень четвертой кратности $y = 0$. Корни $f(x, y) = 0$ имеют вид

$$y = x^{1/2}(1+x)^{1/2}, \quad y = x(1+x)^{1/2}, \quad y = -x(1+x)^{1/2}.$$

В обозначениях теоремы 2 мы имеем $m_1 = 2$, $m_2 = 1$, $m_3 = 1$.

Заметим, что хотя полином последнего примера имеет нуль четырехкратным корнем, возмущения в коэффициентах при степенях y связаны так, что ни одно возмущение корней не имеет порядка $x^{1/4}$. Такого рода связанные возмущения обычны в теории собственных значений.

Теория возмущений для простых собственных значений

5. Рассмотрим две матрицы A и B , удовлетворяющие условию (2.1), и пусть λ_1 — простое собственное значение A . Мы хотим найти соответствующее собственное значение $(A + \varepsilon B)$. Пусть характеристическое уравнение A есть

$$\det(\lambda I - A) \equiv \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \dots + c_0 = 0, \quad (5.1)$$

Тогда характеристическое уравнение $(A + \varepsilon B)$ будет

$$\det(\lambda I - A - \varepsilon B) \equiv \lambda^n + c_{n-1}(\varepsilon)\lambda^{n-1} + c_{n-2}(\varepsilon)\lambda^{n-2} + \dots + c_0(\varepsilon) = 0, \quad (5.2)$$

где $c_r(\varepsilon)$ — полиномы степени $(n - r)$ по ε такие, что

$$c_r(0) = c_r. \quad (5.3)$$

Это становится очевидным, если мы напишем точное выражение для $\det(\lambda I - A - \varepsilon B)$. Мы можем положить

$$c_r(\varepsilon) = c_r + c_{r1}\varepsilon + c_{r2}\varepsilon^2 + \dots + c_{r, n-r}\varepsilon^{n-r}. \quad (5.4)$$

Теперь, так как λ_1 — простой корень (5.1), мы знаем из теоремы 1 § 3, что для достаточно малых ε существует простой корень (5.2), который дается сходящимся степенным рядом

$$\lambda_1(\varepsilon) = \lambda_1 + k_1\varepsilon + k_2\varepsilon^2 + \dots \quad (5.5)$$

Очевидно, что $\lambda_1(\varepsilon) \rightarrow \lambda_1$ при $\varepsilon \rightarrow 0$. Заметим, что

$$|\lambda_1(\varepsilon) - \lambda_1| = O(\varepsilon) \quad (5.6)$$

независимо от кратности остальных собственных значений

Возмущение соответствующих собственных векторов

6. Прежде чем рассматривать возмущение соответствующего собственного вектора, сначала выведем явное выражение для компонент собственного вектора x_1 , соответствующего собственному значению λ_1 матрицы A . Так как λ_1 — простое собственное значение, $(A - \lambda_1 I)$ имеет по крайней мере один отличный от нуля минор порядка $(n - 1)$. Без ограничения общности предположим, что он лежит в первых $(n - 1)$ строках $(A - \lambda_1 I)$. Тогда из теории линейных уравнений известно, что в качестве компонент x можно взять

$$(A_{n1}, A_{n2}, \dots, A_{nn}), \quad (6.1)$$

где A_{ni} — алгебраическое дополнение (n, i) -го элемента $(A - \lambda_1 I)$, и, следовательно, A_{ni} — это полиномы по λ_1 степени не выше $(n - 1)$.

Мы применим этот результат к простому собственному значению $\lambda_1(\varepsilon)$ матрицы $(A + \varepsilon B)$. Обозначим собственный вектор A через x_1 , собственный вектор $(A + \varepsilon B)$ через $x_1(\varepsilon)$. Ясно, что элементы $x_1(\varepsilon)$ — это полиномы по $\lambda_1(\varepsilon)$ и ε , и так как степенной ряд для $\lambda_1(\varepsilon)$ сходится для достаточно малых ε , то каждый элемент $x_1(\varepsilon)$ представим сходящимся степенным рядом по ε , постоянный член в котором — это соответствующий элемент x_1 . Можно написать

$$x_1(\varepsilon) = x_1 + \varepsilon z_1 + \varepsilon^2 z_2 + \dots, \quad (6.2)$$

где каждая компонента векторного ряда в правой части — сходящийся степенной ряд по ε . Аналогично результату (5.6) для собственного значения, получаем для собственного вектора

$$|x_1(\varepsilon) - x_1| = O(\varepsilon), \quad (6.3)$$

и снова здесь нет дробных степеней ε .

Матрица с линейными элементарными делителями

7. Если матрица A имеет линейные элементарные делители, то мы можем предположить, что существуют полные системы правых и левых собственных векторов $x_1, x_2, \dots, x_n$ и $y_1, y_2, \dots, y_n$ таких, что

$$y_i^T x_j = 0 \quad (i \neq j), \quad (7.1)$$

хотя эти векторы единственны, только если все собственные значения простые.

Выразим каждый вектор z_i в (6.2) через x_j в виде

$$z_i = \sum_{j=1}^n s_{ji} x_j. \quad (7.2)$$

Тогда имеем

$$x_1(\varepsilon) = x_1 + \varepsilon \sum_{j=1}^n s_{j1} x_j + \varepsilon^2 \sum_{j=1}^n s_{j2} x_j + \dots, \quad (7.3)$$

и, собирая вместе члены с x_i , получим

$$x_1(\varepsilon) = (1 + \varepsilon s_{11} + \varepsilon^2 s_{12} + \dots) x_1 + (\varepsilon s_{21} + \varepsilon^2 s_{22} + \dots) x_2 + \dots + (\varepsilon s_{n1} + \varepsilon^2 s_{n2} + \dots) x_n. \quad (7.4)$$

Сходимость n степенных рядов, стоящих в скобках, есть простое следствие абсолютной сходимости рядов в (6.2).

Так как нас интересует $x_1(\varepsilon)$ с точностью до множителя, можем разделить обе части (7.4) на $(1 + \varepsilon s_{11} + \varepsilon^2 s_{12} + \dots)$, ибо при достаточно малых ε это не ноль. Мы можем назвать новый вектор снова $x_1(\varepsilon)$. Тогда

$$x_1(\varepsilon) = x_1 + (\varepsilon t_{21} + \varepsilon^2 t_{22} + \dots) x_2 + \dots + (\varepsilon t_{n1} + \varepsilon^2 t_{n2} + \dots) x_n, \quad (7.5)$$

причем выражения в скобках суть также сходящиеся степенные ряды для достаточно малых ε .

Это соотношение получено при предположении, что компоненты x_i выражены прямо через определители. Однако если мы заменим эти x_i новыми x_i , нормированными так, что

$$\|x_i\|_2 = 1, \quad (7.6)$$

то в каждое выражение в скобках войдут постоянные множители, и эти множители можно внести в t_{ij} . Выражение (7.5) справедливо и для нормированных x_i , хотя $x_i(\varepsilon)$, конечно, не обязаны быть нормированными для $\varepsilon \neq 0$.

Возмущения первого порядка собственных значений

8. Сейчас мы выведем точное выражение для возмущения первого порядка в терминах x_i и y_i . Сначала введем величины s_i , которые будут часто использоваться в дальнейшем. Они определяются соотношением

$$s_i = y_i^T x_i \quad (i = 1, 2, \dots, n); \quad (8.1)$$

где y_i и x_i — нормированные левые и правые векторы. Если y_i и x_i вещественны, то s_i есть косинус угла между этими векторами.

Если имеются кратные собственные значения, соответствующие x_i и y_i не будут однозначно определены. В этом случае предполагаем, что s_i , которые мы используем, соответствуют некоторому выбору x_i и y_i . Даже когда x_i и y_i соответствуют простому λ_i , они определены с точностью до комплексного множителя, равного по модулю единице, но в этом случае величина $|s_i|$ будет полностью определена.

В любом случае имеем

$$|s_i| = |y_i^T x_i| \leq \|y_i\|_2 \cdot \|x_i\|_2 = 1. \quad (8.2)$$

Определим также величины β_{ij} соотношениями

$$\beta_{ij} = y_i^T B x_j. \quad (8.3)$$

Так как $\|B\|_2 \leq n$, мы имеем

$$|\beta_{ij}| = |y_i^T B x_j| \leq \|y_i\|_2 \|B x_j\|_2 \leq \|B\|_2 \|y_i\|_2 \|x_j\|_2 \leq n. \quad (8.4)$$

9. По определению

$$(A + \varepsilon B) x_1(\varepsilon) = \lambda_1(\varepsilon) x_1(\varepsilon), \quad (9.1)$$

и так как $\lambda_1(\varepsilon)$ и все компоненты $x_1(\varepsilon)$ представляются сходящимися степенными рядами, мы можем приравнять члены при одинаковых степенях ε в этом уравнении. Приравнявая коэффициенты при ε и используя (5.5) и (7.5), получим

$$A \left(\sum_{i=2}^n t_{i1} x_i \right) + B x_1 = \lambda_1 \left(\sum_{i=2}^n t_{i1} x_i \right) + k_1 x_1, \quad (9.2)$$

или

$$\sum_{i=2}^n (\lambda_i - \lambda_1) t_{i1} x_i + B x_1 = k_1 x_1. \quad (9.3)$$

Умножая слева на y_1^T и помня, что $y_1^T x_i = 0$ ($i \neq 1$), получим

$$k_1 = \frac{y_1^T B x_1}{y_1^T x_1} = \frac{\beta_{11}}{s_1}, \quad (9.4)$$

и, следовательно, из (8.4)

$$|k_1| \leq \frac{n}{|s_1|}. \quad (9.5)$$

При достаточно малом ε главный член в возмущении λ_1 равен $k_1 \varepsilon$, так что чувствительность этого собственного значения в первую очередь зависит от s_1 . К сожалению, s_i могут быть произвольно малыми.

Возмущения первого порядка собственных векторов

10. Умножая (9.3) слева на y_i^T , получим

$$(\lambda_i - \lambda_1) t_{i1} s_i + \beta_{i1} = 0 \quad (i = 2, 3, \dots, n), \quad (10.1)$$

и, следовательно, из (7.5) следует, что член первого порядка в возмущении x_1 имеет вид

$$\varepsilon \left[\frac{\beta_{21} x_2}{(\lambda_1 - \lambda_2) s_2} + \frac{\beta_{31} x_3}{(\lambda_1 - \lambda_3) s_3} + \dots + \frac{\beta_{n1} x_n}{(\lambda_1 - \lambda_n) s_n} \right]. \quad (10.2)$$

Заметим, что здесь участвуют все s_i , кроме s_1 , и мы имеем в знаменателях множители $(\lambda_1 - \lambda_i)$. Присутствие s_i вводит в заблуждение, как мы покажем в § 26. Однако, можно ожидать, что собственный вектор, соответствующий простому собственному значению λ_1 , будет более чувствителен к возмущениям A , если λ_1 *близко* к каким-либо другим собственным значениям, и это действительно так.

Если λ_1 хорошо отделено от остальных собственных значений и ни одно из s_i ($i = 2, 3, \dots, n$) не мало, можем сказать, что собственный вектор x_1 сравнительно нечувствителен к возмущениям A .

Возмущения высоких порядков

11. Мы можем получать выражения для возмущений высших порядков, приравнявая коэффициенты при высших степенях ϵ в (9.1). Например, приравнявая коэффициенты при ϵ^2 , получим

$$A \left(\sum_{i=2}^n t_{i2} x_i \right) + B \left(\sum_{i=2}^n t_{i1} x_i \right) = k_2 x_1 + k_1 \left(\sum_{i=2}^n t_{i1} x_i \right) + \lambda_1 \left(\sum_{i=2}^n t_{i2} x_i \right), \quad (11.1)$$

или

$$\sum_{i=2}^n t_{i2} (\lambda_i - \lambda_1) x_i + B \left(\sum_{i=2}^n t_{i1} x_i \right) = k_2 x_1 + k_1 \left(\sum_{i=2}^n t_{i1} x_i \right).$$

Умножение на y_1^T слева дает

$$\sum_{i=2}^n t_{i1} \beta_{1i} = k_2 s_1, \quad (11.2)$$

и, следовательно, из (10.1)

$$k_2 = \frac{1}{s_1} \sum_{i=2}^n \frac{\beta_{i1} \beta_{1i}}{s_i (\lambda_1 - \lambda_i)}. \quad (11.3)$$

Таким же образом мы можем получить члены второго порядка в возмущении x_1 и т. д., но эти члены высшего порядка имеют малую практическую ценность.

Кратные собственные значения

12. При рассмотрении возмущения собственного значения кратности m анализ, аналогичный проведенному в § 5, требует применения теоремы 2 § 3. В силу этой теоремы такому собственному значению соответствует система m собственных значений, которые, вообще говоря, распадаются в группы и могут быть разложены в дробные степенные ряды по ϵ . Если имеется лишь одна группа, то каждое собственное значение представимо степенным рядом по $\epsilon^{1/m}$, но если имеется несколько групп, то возникают и другие дробные степени. В частности, мы можем иметь m групп, и в этом случае не будет дробных степеней, и для каждого из m возмущенных собственных значений $\lambda_i(\epsilon)$ будет справедливо соотношение вида

$$|\lambda_i(\epsilon) - \lambda_i^*| = O(\epsilon)^{\frac{1}{m}} \quad (12.1)$$

Анализ в терминах характеристического полинома для кратных нулей слишком груб, так как полином недостаточно отражает строение матрицы. В следующих параграфах мы рассмотрим теорию возмущений, использующую значительно более мощные методы, которые имеют также большую практическую важность.

Теоремы Гершгорина

13. При этом новом подходе нам потребуются две теоремы, доказанные Гершгориным (1931).

Т е о р е м а 3. Любое собственное значение матрицы A лежит по крайней мере в одном из кругов с центрами a_{ii} и радиусами $\sum_{j \neq i} |a_{ij}|$.

Доказательство весьма просто. Пусть λ — любое собственное значение A . Тогда есть по крайней мере один ненулевой x такой, что

$$Ax = \lambda x. \quad (13.1)$$

Предположим, что наибольший модуль имеет r -я компонента x . Можно нормировать x так, что

$$x^T = (x_1, x_2, \dots, x_{r-1}, 1, x_{r+1}, \dots, x_n), \quad (13.2)$$

где

$$|x_i| \leq 1 \quad (i \neq r). \quad (13.3)$$

Приравнивая r -е компоненты (13.1), получим

$$\sum_{j=1}^n a_{rj} x_j = \lambda x_r = \lambda. \quad (13.4)$$

Следовательно,

$$|\lambda - a_{rr}| \leq \sum_{j \neq r} |a_{rj} x_j| \leq \sum_{j \neq r} |a_{rj}| |x_j| \leq \sum_{j \neq r} |a_{rj}|, \quad (13.5)$$

и λ лежит в одном из наших кругов.

Вторая теорема дает более детальную информацию относительно распределения собственных значений по кругам.

Т е о р е м а 4. Если s кругов теоремы 3 образуют связную область, изолированную от остальных кругов, то в этой связной области находится ровно s собственных значений матрицы A .

Доказательство основано на понятии непрерывности. Положим

$$A = \text{diag}(a_{ii}) + C = D + C, \quad (13.6)$$

где C — матрица, недиагональные элементы которой равны соответствующим элементам A , а диагональные равны нулю, и определим величины r_i соотношениями

$$r_i = \sum_{j \neq i} |a_{ij}|. \quad (13.7)$$

Рассмотрим теперь матрицы $(D + \varepsilon C)$ для значений ε :

$$0 \leq \varepsilon \leq 1. \quad (13.8)$$

При $\varepsilon = 0$ это матрица D , а при $\varepsilon = 1$ — матрица A . Коэффициенты характеристического полинома $(D + \varepsilon C)$ суть полиномы от ε , и по теории алгебраических функций корни характеристического полинома являются непрерывными функциями ε . По теореме 3 при любом значении ε все собственные значения лежат в кругах с центрами a_{ii} и радиусами εr_i , и если мы меняем ε непрерывно от 0 до 1, все собственные значения меняются непрерывно.

Без ограничения общности мы можем считать, что именно первые s кругов образуют связную область. Тогда, так как остальные $(n - s)$ кругов с радиусами $r_{s+1}, r_{s+2}, \dots, r_n$ изолированы от кругов с радиу-

сами $r_1, r_2, \dots, r_s$, это же верно для соответствующих кругов с радиусами εr_i для всех ε из промежутка $[0, 1)$. Но при $\varepsilon = 0$ собственные значения равны $a_{11}, a_{22}, \dots, a_{nn}$, и первые s из них лежат в области, соответствующей первым s кругам, а остальные лежат вне этой области. Отсюда следует, что это же верно для всех ε , включая $\varepsilon = 1$.

В частности, если какой-либо из кругов Гершгорина изолирован, то он содержит точно одно собственное значение. Заметим, что аналогичный результат можно получить, рассматривая A_ε^T вместо A .

Теория возмущений, основанная на теоремах Гершгорина

14. Мы будем исследовать собственные значения $(A + \varepsilon B)$, используя жорданову каноническую форму. Будем различать пять главных случаев.

С л у ч а й 1. *Возмущение простого собственного значения λ_1 матрицы, имеющей линейные элементарные делители.*

В этом случае жорданова каноническая форма диагональна. Существует матрица H такая, что

$$H^{-1}AH = \text{diag}(\lambda_i); \quad (14.1)$$

причем ее столбцы составляют полную систему правых собственных векторов x_i , а строки H^{-1} составляют полную систему левых собственных векторов y_i^T . Если мы нормируем эти векторы так, что

$$\|x_i\|_2 = \|y_i\|_2 = 1, \quad (14.2)$$

то можем взять в качестве i -го столбца H вектор x_i , и тогда i -я строка H^{-1} будет y_i^T/s_i , в обозначениях § 8. (Заметим, что так как λ_1 — простое собственное значение, x_1 и y_1 единственны с точностью до множителя, равного по модулю единице.) Тогда

$$H^{-1}(A + \varepsilon B)H = \text{diag } \lambda_i + \varepsilon \begin{bmatrix} \beta_{11}/s_1 & \beta_{12}/s_1 & \dots & \beta_{1n}/s_1 \\ \beta_{21}/s_2 & \beta_{22}/s_2 & \dots & \beta_{2n}/s_2 \\ \dots & \dots & \dots & \dots \\ \beta_{n1}/s_n & \beta_{n2}/s_n & \dots & \beta_{nn}/s_n \end{bmatrix}. \quad (14.3)$$

Применяя теорему 3, получим, что собственные значения лежат в кругах с центрами $(\lambda_i + \varepsilon \beta_{ii}/s_i)$ и радиусами $\varepsilon \sum_{j \neq i} |\beta_{ij}/s_i|$. Используя соотношение (8.4), видим, что если $|b_{ij}| < 1$, то радиус i -го круга меньше чем $\frac{n(n-1)\varepsilon}{|s_i|}$, и так как λ_i предполагалось простым, для достаточно малых ε первый круг изолирован и поэтому содержит в точности одно собственное значение.

15. Результат, который мы только что получили, слегка разочаровывает, так как мы могли бы ожидать из предыдущего анализа, что собственное значение, соответствующее λ_1 , должно находиться в круге с центром $(\lambda_1 + \varepsilon \beta_{11}/s_1)$, но имеющем радиус $O(\varepsilon^2)$ при $\varepsilon \rightarrow 0$. Мы можем уменьшать радиусы кругов Гершгорина следующим простым способом.

Если мы умножим i -й столбец какой-либо матрицы на m , а i -ю строку на $1/m$, то ее собственные значения не изменятся. Применим это для $i = 1$ к матрице, стоящей в правой части (14.3), взяв $m = k/\varepsilon$. Тогда она

становится

$$\text{diag}(\lambda_i) + \begin{bmatrix} \varepsilon\beta_{11}/s_1 & \varepsilon^2\beta_{12}/ks_1 & \varepsilon^2\beta_{13}/ks_1 & \dots & \varepsilon^2\beta_{1n}/ks_1 \\ k\beta_{21}/s_2 & \varepsilon\beta_{22}/s_2 & \varepsilon\beta_{23}/s_2 & \dots & \varepsilon\beta_{2n}/s_2 \\ \dots & \dots & \dots & \dots & \dots \\ k\beta_{n1}/s_n & \varepsilon\beta_{n2}/s_n & \varepsilon\beta_{n3}/s_n & \dots & \varepsilon\beta_{nn}/s_n \end{bmatrix}. \quad (15.1)$$

Все элементы в первой строке, кроме элемента в позиции (1.1), теперь содержат множитель ε^2 , в то время как элементы в первом столбце, кроме элемента в позиции (1.1), не зависят от ε . Все остальные элементы не изменились. Мы хотим выбрать k так, чтобы сделать первый круг Гершгорина возможно более малым, сохраняя остальные круги достаточно малыми, чтобы избежать пересечения с первым.

Ясно, что это будет выполнено для всех достаточно малых ε , если мы выберем k наибольшим, удовлетворяющим неравенствам

$$\left| \frac{k\beta_{i1}}{s_i} \right| \leq \frac{1}{2} |\lambda_1 - \lambda_i| \quad (i = 2, 3, \dots, n). \quad (15.2)$$

(Множитель $1/2$ в соотношениях (15.2) не имеет особого значения: его можно заменить любым другим, не зависящим от ε и меньшим единицы). Это требование будет выполнено, если

$$k = \min \left| \frac{(\lambda_1 - \lambda_i) s_i}{2\beta_{i1}} \right| \quad (i = 2, 3, \dots, n). \quad (15.3)$$

(Если все β_{i1} — нули, то $(\lambda_1 + \varepsilon\beta_{11}/s_1)$ — точное собственное значение, и нам не надо более обсуждать этот случай.) При таком определении k имеем для радиуса первого круга Гершгорина

$$r_1 = \sum_{j=2}^n \left| \frac{\varepsilon^2\beta_{1j}}{ks_1} \right| \leq \frac{n(n-1)\varepsilon^2}{k|s_1|}. \quad (15.4)$$

Мы также имеем

$$k^{-1} = \max \left[\frac{2\beta_{i1}}{(\lambda_1 - \lambda_i) s_i} \right] \leq \max \left[\frac{2n}{(\lambda_1 - \lambda_i) s_i} \right]. \quad (15.5)$$

Напомним читателю, что эти оценки получены при предположении, что B нормирована так, что $|b_{ij}| \leq 1$.

16. Явные выражения для границ, вообще говоря, скрывают простоту основной техники. Эта простота становится ясной при рассмотрении численных примеров. Для иллюстрации рассмотрим сначала матрицу

$$X = \begin{bmatrix} 0,9 & & \\ & 0,4 & \\ & & 0,4 \end{bmatrix} + 10^{-5} \begin{bmatrix} 0,1234 & 0,4132 & -0,2167 \\ -0,1342 & 0,4631 & 0,1276 \\ 0,1567 & 0,1432 & 0,3125 \end{bmatrix} \quad (16.1)$$

Первая матрица справа имеет одно двукратное собственное значение но линейные элементарные делители.

Умножая первую строку на 10^{-5} и первый столбец на 10^5 , видим, что собственные значения X совпадают с собственными значениями матрицы

$$\begin{bmatrix} 0,9 & & \\ & 0,4 & \\ & & 0,4 \end{bmatrix} + \begin{bmatrix} 10^{-5}(0,1234) & 10^{-10}(0,4132) & 10^{-10}(-0,2167) \\ -0,1342 & 10^{-5}(0,4631) & 10^{-5}(0,1276) \\ 0,1567 & 10^{-5}(0,1432) & 10^{-5}(0,3125) \end{bmatrix}. \quad 16.2$$

Соответствующие круги Гершгорина имеют:

центр $0,9 + 10^{-5}(0,1234)$, радиус $10^{-10}(0,4132 + 0,2167) = 10^{-10}(0,6299)$;

центр $0,4 + 10^{-5}(0,4631)$, радиус $0,1342 + 10^{-5}(0,1276)$;

центр $0,4 + 10^{-5}(0,3125)$, радиус $0,1567 + 10^{-5}(0,1432)$.

Первый круг изолирован и поэтому содержит в точности одно собственное значение; радиус этого круга порядка 10^{-10} . Тот факт, что другие два диагональных элемента совпадают, не влияет на результат в целом.

Однако если мы позволим одному из остальных двух собственных значений приблизиться к первому, то отделение этого собственного значения будет затруднено. Рассмотрим матрицу Y :

$$Y = \begin{bmatrix} 0,9 & & \\ & 0,89 & \\ & & 0,4 \end{bmatrix} + 10^{-5} \begin{bmatrix} 0,1234 & 0,4132 & -0,2167 \\ -0,1342 & 0,4631 & 0,1276 \\ 0,1567 & 0,1432 & 0,3125 \end{bmatrix}. \quad (16.3)$$

Мы не можем использовать теперь множитель 10^{-5} , так как тогда бы первые два круга Гершгорина пересеклись. Используя множитель 10^{-3} , получаем

$$\begin{bmatrix} 0,9 & & \\ & 0,89 & \\ & & 0,4 \end{bmatrix} + \begin{bmatrix} 10^{-5} & (0,1234) & 10^{-8} & (0,4132) & 10^{-8} & (-0,2167) \\ 10^{-2} & (-0,1342) & 10^{-5} & (0,4631) & 10^{-5} & (0,1276) \\ 10^{-2} & (0,1567) & 10^{-5} & (0,1432) & 10^{-5} & (0,3125) \end{bmatrix}, \quad (16.4)$$

и первые два круга Гершгорина изолированы. Следовательно, существует собственное значение в круге с центром $0,9 + 10^{-5}(0,1234)$ и радиусом $10^{-8}(0,6299)$.

17. С л у ч а й 2. Возмущение кратного собственного значения λ_1 матрицы, имеющей линейные элементарные делители.

Этот случай вполне иллюстрируется простым примером матрицы шестого порядка, у которой $\lambda_1 = \lambda_2 = \lambda_3$ и $\lambda_4 = \lambda_5$. Так как эта матрица A имеет линейные делители, $A + \varepsilon B$ подобна

$$\begin{bmatrix} \lambda_1 & & & & & \\ & \lambda_1 & & & & \\ & & \lambda_1 & & & \\ & & & \lambda_4 & & \\ & & & & \lambda_4 & \\ & & & & & \lambda_6 \end{bmatrix} + \varepsilon \begin{bmatrix} \beta_{11}/s_1 & \beta_{12}/s_1 & \dots & \beta_{16}/s_1 \\ \beta_{21}/s_2 & \beta_{22}/s_2 & \dots & \beta_{26}/s_2 \\ \dots & \dots & \dots & \dots \\ \beta_{61}/s_6 & \beta_{62}/s_6 & \dots & \beta_{66}/s_6 \end{bmatrix}. \quad (17.1)$$

Мы уже имели дело с λ_6 в случае 1. Умножая шестой столбец на k/ε и шестую строку на ε/k , мы можем локализовать простое собственное значение в круге с центром $\lambda_6 + \varepsilon\beta_{66}/s_6$ и радиусом, имеющим порядок $O(\varepsilon^2)$ при $\varepsilon \rightarrow 0$. Рассмотрим теперь остальные пять кругов Гершгорина. Три из них имеют центры в $\lambda_1 + \varepsilon\beta_{ii}/s_i$ ($i = 1, 2, 3$) и два — в $\lambda_4 + \varepsilon\beta_{ii}/s_i$ ($i = 4, 5$), и все соответствующие радиусы порядка ε . Очевидно, что при достаточно малых ε группа из трех кругов будет изолирована от группы из двух, но мы не можем, вообще говоря, утверждать, что отдельные круги в группе будут изолированы. Все возмущения трехкратного собственного значения λ_1 имеют порядок ε при $\varepsilon \rightarrow 0$, но, вообще говоря, неверно, что существует по одному собственному значению в каждом из трех кругов с центрами $\lambda_1 + \varepsilon\beta_{ii}/s_i$ ($i = 1, 2, 3$) и радиусами порядка ε^2 .

18. Мы можем, однако, несколько уменьшить радиусы этих кругов. Остановимся на трехкратном собственном значении. Перепишем (17.1) в виде

$$\text{diag}(\lambda_i) + \varepsilon \begin{bmatrix} P & Q \\ R & S \end{bmatrix}, \quad (18.1)$$

где P, Q, R, S — матрицы третьего порядка. Умножая первые три строки на ε/k и первые три столбца на k/ε , получим

$$\text{diag}(\lambda_i) + \begin{bmatrix} \varepsilon P & k^{-1}\varepsilon^2 Q \\ kR & \varepsilon S \end{bmatrix}. \quad (18.2)$$

Мы можем выбрать значение k независимо от ε так, что первые три круга станут изолированными от остальных. Следовательно, при соответствующих k три собственных значения находятся в объединении трех кругов:

$$\begin{aligned} \text{центр } \lambda_1 + \frac{\varepsilon\beta_{11}}{s_1}, \quad \text{радиус } \frac{\varepsilon(|\beta_{12}| + |\beta_{13}|)}{|s_1|} + \varepsilon^2 \frac{|\beta_{14}| + |\beta_{15}| + |\beta_{16}|}{k|s_1|}, \\ \text{центр } \lambda_1 + \frac{\varepsilon\beta_{22}}{s_2}, \quad \text{радиус } \frac{\varepsilon(|\beta_{21}| + |\beta_{23}|)}{|s_2|} + \varepsilon^2 \frac{|\beta_{24}| + |\beta_{25}| + |\beta_{26}|}{k|s_2|}, \\ \text{центр } \lambda_1 + \frac{\varepsilon\beta_{33}}{s_3}, \quad \text{радиус } \frac{\varepsilon(|\beta_{31}| + |\beta_{32}|)}{|s_3|} + \varepsilon^2 \frac{|\beta_{34}| + |\beta_{35}| + |\beta_{36}|}{k|s_3|}. \end{aligned}$$

Для матрицы шестого порядка это уменьшение радиусов незаметно, но для матриц высших порядков оно существенно. Мы увидим далее (гл. 9, § 68), что, приводя матрицу P в (18.1) к канонической форме, можно получать более точные результаты, но здесь мы не останавливаемся на этом.

Мы показали, что если матрица A имеет кратные собственные значения, но линейные элементарные делители, то коэффициенты характеристического полинома должны быть связаны таким образом, что в возмущениях не возникают дробные степени ε .

19. С л у ч а й 3. *Возмущение простого собственного значения матрицы, имеющей один или более нелинейных элементарных делителей.*

Так как A имеет нелинейные элементарные делители, то уже не существует полной системы собственных векторов. Сначала исследуем величины, играющие роль s_i для таких матриц. Рассмотрим простую матрицу A вида

$$A = \begin{bmatrix} a & 1 \\ 0 & b \end{bmatrix}. \quad (19.1)$$

Собственные значения этой матрицы $\lambda_1 = a$ и $\lambda_2 = b$, и соответствующие правые и левые векторы суть

$$\left. \begin{aligned} x_1^T &= (1, 0), & \alpha x_2^T &= (1, b - a), \\ \alpha y_1^T &= (a - b, 1), & y_2^T &= (0, 1), \\ \alpha &= [1 + (a - b)^2]^{1/2}. \end{aligned} \right\} \quad (19.2)$$

Поэтому имеем

$$s_1 = y_1^T x_1 = \frac{a - b}{\alpha}, \quad s_2 = y_2^T x_2 = \frac{b - a}{\alpha}, \quad (19.3)$$

и при $b \rightarrow a$ оба s_1 и s_2 стремятся к нулю. Мы видели, что чувствительность простого собственного значения к возмущениям пропорциональна s_i^{-1} , и, следовательно, когда b стремится к a , оба собственных значения становятся все более и более чувствительными. Заметим, однако, что мы имеем $s_1^{-1} + s_2^{-1} = 0$ при всех значениях b , так что хотя s_1^{-1} и s_2^{-1} стремятся к бесконечности, они не независимы.

20. Для матриц с нелинейными элементарными делителями анализ, аналогичный проведенному в предыдущих параграфах, может быть проведен с использованием жордановой канонической формы A . Существует неособенная матрица H такая, что

$$H^{-1}AH = C, \quad (20.1)$$

где C — верхняя каноническая жорданова матрица. Мы теперь не можем предполагать, что все столбцы H нормированы, так как масштабные множители столбцов, связанных с нелинейными элементарными делителями, подчинены требованию, чтобы соответствующие наддиагональные элементы C были единицы (гл. 1, § 8). Тем не менее все же удобно предполагать, что столбцы, связанные с линейными элементарными делителями, нормированы. Обозначим через G матрицу, состоящую из нормированных строк H^{-1} , и будем писать

$g_i^T h_i = s_i$, если g_i и h_i соответствуют линейным делителям;

$g_i^T h_i = t_i$, если g_i и h_i являются частью системы, соответствующей нелинейному элементарному делителю.

Возмущение простого собственного значения вполне иллюстрируется на простом примере. Рассмотрим матрицу четвертого порядка, имеющую элементарные делители $(\lambda_1 - \lambda)^2$, $(\lambda_3 - \lambda)$, $(\lambda_4 - \lambda)$. Можно написать

$$H^{-1}(A + \varepsilon B)H = \begin{bmatrix} \lambda_1 & 1 & & \\ 0 & \lambda_1 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{bmatrix} + \varepsilon \begin{bmatrix} \beta_{11}/t_1 & \beta_{12}/t_1 & \beta_{13}/t_1 & \beta_{14}/t_1 \\ \beta_{21}/t_2 & \beta_{22}/t_2 & \beta_{23}/t_2 & \beta_{24}/t_2 \\ \beta_{31}/s_3 & \beta_{32}/s_3 & \beta_{33}/s_3 & \beta_{34}/s_3 \\ \beta_{41}/s_4 & \beta_{42}/s_4 & \beta_{43}/s_4 & \beta_{44}/s_4 \end{bmatrix}, \quad (20.2)$$

где

$$\beta_{ij} = g_i^T B h_j \quad (\text{все } i, j), \quad |\beta_{ij}| \leq n \quad (i, j \geq 3). \quad (20.3)$$

Естественно задать вопрос, будут ли возмущения простых собственных значений λ_3 и λ_4 такими, что

$$\left. \begin{aligned} |\lambda_3(\varepsilon) - \lambda_3| &= O(\varepsilon) \\ |\lambda_4(\varepsilon) - \lambda_4| &= O(\varepsilon) \end{aligned} \right\} \quad \text{при} \quad \varepsilon \rightarrow 0. \quad (20.4)$$

На первый взгляд присутствие единицы в жордановой форме является большим неудобством, так как радиус первого круга Гершгорина больше единицы при всех ε . Следовательно, если $|\lambda_1 - \lambda_3|$ меньше единицы, третий круг не изолирован от первого при всех ε . Однако если мы умножим первую строку (20.2) на m и первый столбец на $1/m$, то получим

$$\begin{bmatrix} \lambda_1 & m & & \\ 0 & \lambda_1 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{bmatrix} + \varepsilon \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{bmatrix}, \quad (20.5)$$

где γ_{ij} не зависят от ε . Выбор

$$m = \min \left\{ \frac{1}{4} |\lambda_1 - \lambda_3|, \frac{1}{4} |\lambda_1 - \lambda_4| \right\} \quad (20.6)$$

устраняет трудность. Чтобы изолировать третий круг Гершгорина и сделать его радиус порядка ε^2 , умножим третий столбец и третью строку на k/ε и ε/k соответственно, выбрав k так, чтобы

$$|k\gamma_{i3}| \leq \frac{1}{2} |\lambda_3 - \lambda_i| \quad (i \neq 3). \quad (20.7)$$

Точно так же, как в § 15, мы видим, что круг с центром $\lambda_3 + \varepsilon\gamma_{33}$ и радиусом $\varepsilon^2 (|\gamma_{31}| + |\gamma_{32}| + |\gamma_{34}|)$ изолирован при достаточно малых ε и поэтому содержит в точности одно собственное значение. Следовательно, возмущение простого собственного значения λ_3 таково, что

$$|\lambda_3(\varepsilon) - \lambda_3 - \varepsilon\gamma_{33}| = O(\varepsilon^2) \quad \text{при } \varepsilon \rightarrow 0. \quad (20.8)$$

Здесь γ_{33} есть на самом деле β_{33}/s_3 и из (20.2) мы имеем $|\beta_{33}| \leq n$. Присутствие элементарного делителя $(\lambda_1 - \lambda)^2$ не вызывает существенного различия в поведении простого собственного значения, отличного от λ_1 . Читатель может легко убедиться в том, что этот результат общий.

21. С л у ч а й 4. Возмущение собственного значения, соответствующего нелинейному элементарному делителю полной матрицы.

Рассмотрим нелинейный делитель $(\lambda_1 - \lambda)^2$, и так как предполагается, что матрица полная, то не существует других делителей с множителем $(\lambda_1 - \lambda)$. Матрица § 20 служит иллюстрацией этого случая. Мы рассмотрим возмущения двукратного собственного значения $\lambda = \lambda_1$. В (20.2) элемент, равный единице, стоит в одной из строк, содержащих элемент λ_1 . Если мы хотим получить круг Гершгорина с радиусом, стремящимся к нулю при $\varepsilon \rightarrow 0$, то этот элемент надо умножить на некоторый множитель. На самом деле, умножая второй столбец на $\varepsilon^{1/2}$ и вторую строку на $\varepsilon^{-1/2}$, получаем

$$\begin{bmatrix} \lambda_1 & \varepsilon^{1/2} & & \\ 0 & \lambda_1 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{bmatrix} + \begin{bmatrix} \varepsilon & \varepsilon^{3/2} & \varepsilon & \varepsilon \\ \varepsilon^{1/2} & \varepsilon & \varepsilon^{1/2} & \varepsilon^{1/2} \\ \varepsilon & \varepsilon^{3/2} & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon^{3/2} & \varepsilon & \varepsilon \end{bmatrix}, \quad (21.1)$$

где мы опустили постоянные множители в элементах второй матрицы. Ясно, что члены с $\varepsilon^{1/2}$ преобладают в формуле для радиусов кругов Гершгорина. При достаточно малых ε существуют в круге с центром λ_1 и радиусом порядка $\varepsilon^{1/2}$ два собственных значения. Аналогично для кубического делителя мы можем, например, преобразовать матричную сумму

$$\begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \\ & & & \lambda_4 \end{bmatrix} + \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon \end{bmatrix} \quad (21.2)$$

к виду

$$\begin{bmatrix} \lambda_1 & \varepsilon^{1/3} & 0 \\ 0 & \lambda_1 & \varepsilon^{1/3} \\ 0 & 0 & \lambda_1 \\ & & & \lambda_4 \end{bmatrix} + \begin{bmatrix} \varepsilon & \varepsilon^{4/3} & \varepsilon^{5/3} & \varepsilon \\ \varepsilon^{2/3} & \varepsilon & \varepsilon^{4/3} & \varepsilon^{2/3} \\ \varepsilon^{1/3} & \varepsilon^{2/3} & \varepsilon & \varepsilon^{1/3} \\ \varepsilon & \varepsilon^{4/3} & \varepsilon^{5/3} & \varepsilon \end{bmatrix}, \quad (21.3)$$

умножая второй и третий столбцы на $\varepsilon^{1/3}$ и $\varepsilon^{2/3}$ и вторую и третью строки на $\varepsilon^{-1/3}$ и $\varepsilon^{-2/3}$. Теорема Гершгорина тогда дает, что при достаточно малых ε в круге с центром λ_1 и радиусом, пропорциональным $\varepsilon^{1/3}$, находятся три собственных значения.

Общий результат для делителя n -й степени теперь очевиден, и пример, данный в § 2, показывает, что мы не можем избежать множителя $\varepsilon^{1/n}$, если B — матрица общего вида. Конечно, могут быть специальные возмущения порядка ε , для которых возмущения собственных значений будут порядка ε или даже нули.

22. С л у ч а й 5. *Возмущения собственных значений λ_i , когда существует более чем один делитель с множителем $(\lambda_i - \lambda)$ и по крайней мере один из них нелинейный.*

Прежде чем дать общие результаты для этого случая, рассмотрим простой пример матрицы, имеющей жорданову каноническую форму

$$\left| \begin{array}{ccc|cc} \lambda_1 & 1 & 0 & & \\ 0 & \lambda_1 & 1 & & \\ 0 & 0 & \lambda_1 & & \\ \hline & & & \lambda_1 & 1 \\ & & & 0 & \lambda_1 \end{array} \right|. \quad (22.1)$$

Естественно спросить, будут ли здесь все еще два собственных значения в круге с центром λ_1 и радиусом $O(\varepsilon^{1/2})$ при $\varepsilon \rightarrow 0$. Очевидно, нет, так как если мы возмутим элементы в позициях (3,4) и (5,1), то характеристическое уравнение будет

$$(\lambda_1 - \lambda)^5 + \varepsilon^2 = 0, \quad (22.2)$$

так что все возмущения пропорциональны $\varepsilon^{2/5}$. Мы можем доказать, однако, что существует по крайней мере одно возмущение, которое не имеет порядка больше $\varepsilon^{2/5}$. Действительно, рассмотрим все возможные возмущения порядка ε элементов матрицы (22.1). Мы можем разложить характеристическое уравнение по степеням $(\lambda_1 - \lambda)$, и из рассмотрения определителей, входящих в разложение, очевидно, что свободный член имеет вид

$$A_2\varepsilon^2 + A_3\varepsilon^3 + A_4\varepsilon^4 + A_5\varepsilon^5. \quad (22.3)$$

Здесь нет члена с ε или не зависящего от ε . Если корни характеристического уравнения даются

$$\lambda_1 - \lambda_i = p_i, \quad (22.4)$$

то имеем

$$\prod_{i=1}^5 p_i = A_2\varepsilon^2 + A_3\varepsilon^3 + A_4\varepsilon^4 + A_5\varepsilon^5. \quad (22.5)$$

Следовательно, невозможно всем p_i быть большего порядка чем $\varepsilon^{2/5}$. Добавление собственных значений любой кратности, но отличных от λ_1 , очевидно, не влияет на результат.

Возмущения, соответствующие общему распределению нелинейных делителей

23. Присутствие элементарных делителей высшей степени $(\lambda_1 - \lambda)$, следовательно, влияет на чувствительность собственных значений, соответствующих линейным делителям $(\lambda_1 - \lambda)$, но не влияет на собственные значения, отличные от λ_1 . Оставляем для доказательства в качестве упражнения следующие два результата.

Пусть элементарные делители, соответствующие λ_1 , будут

$$(\lambda_1 - \lambda)^{r_1}, (\lambda_1 - \lambda)^{r_2}, \dots, (\lambda_1 - \lambda)^{r_s}, \quad (23.1)$$

где

$$r_1 \geq r_2 \geq \dots \geq r_s, \quad \sum r_i = t. \quad (23.2)$$

Тогда при достаточно малых ε возмущения $p_1, p_2, \dots, p_t$ удовлетворяют соотношению

$$\prod_{i=1}^t p_i = A_s \varepsilon^s + A_{s+1} \varepsilon^{s+1} + \dots + A_t \varepsilon^t, \quad (23.3)$$

и, следовательно, по крайней мере одно λ лежит в круге с центром λ_1 и радиусом $K_1 \varepsilon^{s/t}$ при некотором значении K_1 . С другой стороны, все соответствующие возмущенные собственные значения лежат в круге с центром λ_1 и радиусом $K_2 \varepsilon^{1/r_1}$ при некотором K_2 .

Теория возмущений для собственных векторов матриц, основанная на исследовании жордановой канонической формы

24. Рассматривая жорданову каноническую форму, можно изучать также и возмущения собственных векторов. Мы не будем рассматривать проблему в полной общности, а ограничимся случаем, когда A имеет линейные элементарные делители. Имеем тогда

$$H^{-1}AH = \text{diag}(\lambda_i), \quad (24.1)$$

и столбцы H образуют полную систему собственных векторов $x_1, x_2, \dots, x_n$ матрицы A . Обозначим «соответствующие» собственные векторы матрицы $(A + \varepsilon B)$ через $x_1(\varepsilon), x_2(\varepsilon), \dots, x_n(\varepsilon)$ и собственные векторы $H^{-1}(A + \varepsilon B)H$ через $z_1(\varepsilon), z_2(\varepsilon), \dots, z_n(\varepsilon)$, так что

$$x_i(\varepsilon) = Hz_i(\varepsilon). \quad (24.2)$$

Так как x_i образуют полную систему собственных векторов, имеем

$$Hz_i(\varepsilon) = \alpha_1(\varepsilon)x_1 + \alpha_2(\varepsilon)x_2 + \dots + \alpha_n(\varepsilon)x_n \quad (24.3)$$

при некоторых $\alpha_i(\varepsilon)$.

Для простоты рассмотрим матрицу шестого порядка § 17. Имеем

$$H^{-1}(A + \varepsilon B)H = \begin{bmatrix} \lambda_1 & & & & & \\ & \lambda_1 & & & & \\ & & \lambda_1 & & & \\ & & & \lambda_4 & & \\ & & & & \lambda_4 & \\ & & & & & \lambda_6 \end{bmatrix} + \varepsilon \begin{bmatrix} \beta_{11}/s_1 & \beta_{12}/s_1 & \beta_{13}/s_1 & \dots & \beta_{16}/s_1 \\ \beta_{21}/s_2 & \beta_{22}/s_2 & \beta_{23}/s_2 & \dots & \beta_{26}/s_2 \\ \dots & \dots & \dots & \dots & \dots \\ \beta_{61}/s_6 & \beta_{62}/s_6 & \beta_{63}/s_6 & \dots & \beta_{66}/s_6 \end{bmatrix}. \quad (24.4)$$

Рассматривая теперь простое собственное значение λ_6 , видим, что $z_6(0) = e_6$. Пусть соответствующий $z_6(\varepsilon)$ нормирован так, что его наибольшая компонента равна единице. При достаточно малых ε это должна быть шестая компонента. Действительно, предположим, что это, например, четвертая, тогда, положив $z_{46}(\varepsilon) = 1$ в соотношении

$$\lambda_6(\varepsilon) [z_{46}(\varepsilon)] = \lambda_4 z_{46}(\varepsilon) + \frac{\varepsilon}{s_4} \sum_{i=1}^6 \beta_{4i} z_{i6}(\varepsilon), \quad (24.5)$$

получим

$$\lambda_6(\varepsilon) - \lambda_4 = \frac{\varepsilon}{s_4} \sum_{i=1}^6 \beta_{4i} z_{i6}(\varepsilon). \quad (24.6)$$

Теперь, если $\varepsilon \rightarrow 0$, то левая часть стремится к $\lambda_6 - \lambda_4$, а правая к нулю. Следовательно, мы имеем противоречие, и поэтому $z_{66}(\varepsilon) = 1$ при достаточно малых ε .

Сейчас покажем, что все остальные компоненты меньше, чем $K\varepsilon$ при $\varepsilon \rightarrow 0$ и некотором K . Действительно,

$$\lambda_6(\varepsilon) z_{i6}(\varepsilon) = \lambda_i z_{i6}(\varepsilon) + \frac{\varepsilon}{s_i} \sum_{j=1}^6 \beta_{ij} z_{j6}(\varepsilon), \quad (24.7)$$

и, следовательно,

$$|\lambda_6(\varepsilon) - \lambda_i| |z_{i6}(\varepsilon)| \leq \frac{\varepsilon}{|s_i|} \sum_{j=1}^6 |\beta_{ij}|, \quad (24.8)$$

что дает

$$|z_{i6}(\varepsilon)| \leq \frac{2\varepsilon \sum_{j=1}^6 |\beta_{ij}|}{|s_i| |\lambda_6 - \lambda_i|} \quad (i = 1, \dots, 5) \quad (24.9)$$

при

$$|\lambda_6(\varepsilon) - \lambda_i| \geq \frac{1}{2} |\lambda_6 - \lambda_i|.$$

Зная, что $z_{i6}(\varepsilon)$ ($i = 1, \dots, 5$) порядка ε , можем получить исправленную границу. На самом деле (24.7) дает

$$(\lambda_6(\varepsilon) - \lambda_i) z_{i6}(\varepsilon) = \frac{\varepsilon \beta_{i6}}{s_i} + \frac{\varepsilon}{s_i} \sum_{j=1}^5 \beta_{ij} z_{j6}(\varepsilon), \quad (24.10)$$

и второй член справа порядка ε^2 . Поэтому имеем

$$\left| z_{i6}(\varepsilon) - \frac{\varepsilon \beta_{i6}}{s_i (\lambda_6 - \lambda_i)} \right| = O(\varepsilon^2). \quad (24.11)$$

Результат в основном совпадает с (10.2), но мы можем теперь видеть, как получать точную границу для члена с ε^2 , и это важно для численных примеров.

**Возмущения собственных векторов, соответствующих
кратным собственным значениям
(линейные элементарные делители)**

25. Мы не можем доказать так много для собственных векторов, соответствующих кратным собственным значениям. Действительно, если $z_5(\varepsilon)$ — нормированный собственный вектор $H^{-1}(A + \varepsilon B)H$, соответствующий $\lambda_5(\varepsilon)$, мы можем показать, как в § 24, что наибольшим элементом $z_5(\varepsilon)$ не может быть ни один из элементов с номерами 1, 2, 3 или 6, и ясно, что эти элементы имеют порядок ε . Нормированный $z_5(\varepsilon)$ должен поэтому быть или вида

$$[z_{15}(\varepsilon), z_{25}(\varepsilon), z_{35}(\varepsilon), k(\varepsilon), 1, z_{65}(\varepsilon)], \quad (25.1)$$

или вида

$$[z_{15}(\varepsilon), z_{25}(\varepsilon), z_{35}(\varepsilon), 1, k(\varepsilon), z_{65}(\varepsilon)], \quad (25.2)$$

где $|k(\varepsilon)| \leq 1$. Соответствующий собственный вектор $x_3(\varepsilon)$ матрицы $(A + \varepsilon B)$ имеет компоненты порядка ε в направлениях x_1, x_2, x_3 и x_6 , но компоненты в подпространстве, образованном x_4 и x_5 , имеют вид либо $x_4 + k(\varepsilon)x_5$, либо $k(\varepsilon)x_4 + x_5$. Мы не можем ожидать улучшения этого результата, так как любой вектор в этом подпространстве является, конечно, собственным вектором A .

Ограничения теории возмущений

26. Несмотря на то, что теория возмущений, развитая нами в §§ 14—25, имеет большое практическое значение, целесообразно на данной стадии обратить внимание на некоторые ее недостатки. Мы изучали природу возмущений при $\varepsilon \rightarrow 0$. Рассмотрим простую матрицу

$$\begin{bmatrix} a & 10^{-10} \\ 0 & a \end{bmatrix}. \quad (26.1)$$

Эта матрица имеет нелинейный элементарный делитель $(a - \lambda)^2$, и если мы добавим возмущение ε к элементу в позиции (2,1), то собственные значения станут $a \pm (10^{-10}\varepsilon)^{1/2}$. Производные по ε этих собственных значений поэтому стремятся к бесконечности при ε , стремящемся к нулю. Однако, если мы интересуемся возмущениями порядка 10^{-10} , этого можно избежать. Если возмущающая матрица имеет вид

$$\begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ \varepsilon_3 & \varepsilon_4 \end{bmatrix}, \quad (26.2)$$

гораздо естественнее рассматривать эту проблему в виде возмущения

$$\begin{bmatrix} 0 & 10^{-10} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ \varepsilon_3 & \varepsilon_4 \end{bmatrix} \quad (26.3)$$

матрицы

$$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}, \quad (26.4)$$

и эта последняя матрица уже не имеет нелинейного элементарного делителя.

Рассматривая собственные векторы, мы раскладывали их возмущения на компоненты в направлениях $x_1, x_2, \dots, x_n$. Когда x_i ортогональны,

это вполне удобно, но если некоторые из x_i «почти» линейно зависимы, то мы можем получить большие компоненты в направлении отдельных x_i , даже если вектор возмущения сам по себе мал. Мы видели в § 10, что возмущение, соответствующее простому собственному значению λ_1 , имеет вид

$$\varepsilon \sum_{i=2}^n \frac{\beta_{i1}}{s_i(\lambda_1 - \lambda_i)} x_i + O(\varepsilon^2), \quad (26.5)$$

и, следовательно, можем ожидать, что, вообще говоря, если s_i мал, то компонента в направлении x_i велика. Однако, как это может ввести в заблуждение, показывает пример матрицы

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 + 10^{-10} \end{bmatrix}. \quad (26.6)$$

Легко показать, что

$$s_1 = 1, \quad s_2 = -10^{-10}/(1 + 10^{-20})^{1/2}, \quad s_3 = 10^{-10}/(1 + 10^{-20})^{1/2}. \quad (26.7)$$

Однако вектор x_1 , соответствующий $\lambda = 2$, без сомнения нечувствителен к возмущениям матрицы. Два же вектора x_2 и x_3 почти одинаковы (так же, как и y_2 и y_3), а множители $\frac{\beta_{21}}{s_2(\lambda_1 - \lambda_2)}$ и $\frac{\beta_{31}}{s_3(\lambda_1 - \lambda_3)}$ почти одинаковы по модулю и противоположны по знаку. Поэтому возмущения в направлениях x_2 и x_3 почти погашаются.

Соотношения между s_i

27. В этом примере $1/s_2$ и $1/s_3$ оба весьма велики, но равны по модулю и имеют разные знаки. Мы сейчас покажем, что s_i взаимно зависят таким образом, что исключается возможность быть большим лишь одному $1/s_i$. Если мы положим

$$x_i = \sum_{j=1}^n \alpha_{ij} y_j, \quad y_i = \sum_{j=1}^n \beta_{ij} x_j, \quad \text{где} \quad \|x_j\|_2 = \|y_j\|_2 = 1, \quad (27.1)$$

то имеем

$$\alpha_{ij} = \frac{x_j^T x_i}{x_j^T y_j} = \frac{x_j^T x_i}{s_j}, \quad \beta_{ij} = \frac{y_j^T y_i}{y_j^T x_j} = \frac{y_j^T y_i}{s_j}, \quad (27.2)$$

$$y_i^T x_i = \sum_{j=1}^n \beta_{ij} x_j^T \sum_{j=1}^n \alpha_{ij} y_j = \sum_{j=1}^n \beta_{ij} \alpha_{ij} s_j = \sum_{j=1}^n (x_j^T x_i) (y_j^T y_i) s_j^{-1}, \quad (27.3)$$

что дает

$$s_i = s_i^{-1} + \sum_{j \neq i} (\cos \theta_{ij} \cos \varphi_{ij}) s_j^{-1}. \quad (27.4)$$

Здесь θ_{ij} — угол между x_i и x_j и φ_{ij} — угол между y_i и y_j . Соотношение (27.4) справедливо при любом i . Из него мы выводим, что

$$|s_i|^{-1} \leq |s_i| + \sum_{j \neq i} |(\cos \theta_{ij} \cos \varphi_{ij}) s_j^{-1}| \leq 1 + \sum_{j \neq i} |s_j|^{-1}. \quad (27.5)$$

Это показывает, например, что мы не можем иметь для матрицы третьего порядка следующего набора s_i :

$$|s_1^{-1}| = 10, \quad |s_2^{-1}| = 10^7, \quad |s_3^{-1}| = 10^8,$$

так как

$$|s_3^{-1}| \gg 1 + |s_1^{-1}| + |s_2^{-1}|.$$

Обусловленность вычислительных проблем

28. Будем говорить, что вычислительная проблема плохо обусловлена, если величины, которые надо вычислить, сильно чувствительны к малым изменениям данных. Рассмотрения предыдущих параграфов выявили многие факторы, от которых зависит чувствительность системы собственных значений и собственных векторов матрицы. Очевидно, что матрица может иметь некоторые собственные значения, которые очень чувствительны к малым изменениям ее элементов, в то время как остальные сравнительно нечувствительны. Аналогично некоторые из собственных векторов могут быть хорошо обусловленными, в то время как другие плохо обусловлены, и собственный вектор может быть плохо обусловлен, а соответствующее собственное значение нет.

Признаки, определяющие, будет ли матрица плохо обусловлена по отношению к решению проблемы собственных значений, отличны от тех, что определяют обусловленность матрицы по отношению к вычислению ее обратной. Как мы увидим в § 11 гл. 4, нормированная матрица должна, конечно, рассматриваться как плохо обусловленная в последнем контексте, если она имеет очень малое собственное значение, но это неверно по отношению к обусловленности проблемы собственных значений. Подходящим выбором k можем сделать $(A - kI)$ особенной, но чувствительность системы собственных значений $(A - kI)$ такая же, как и у A . Однако мы покажем, что существует некоторая связь между обусловленностью этих двух вычислительных проблем.

Числа обусловленности

29. Удобно иметь некоторые числа, определяющие обусловленность матриц по отношению к вычислительным проблемам, и называть такие числа «числами обусловленности». В идеале они должны были бы давать «общую оценку» скорости изменения решения по отношению к изменению в коэффициентах и, следовательно, должны были бы быть каким-либо образом пропорциональны этой скорости изменения.

Из того, что мы говорили в § 28, очевидно, что число обусловленности, даже если мы ограничимся лишь проблемой вычисления собственных значений, должно удовлетворять большим требованиям. Если какое-либо одно собственное значение очень чувствительно, то число обусловленности должно быть большим, даже если остальные собственные значения весьма нечувствительны. Несмотря на эти очевидные ограничения, единое число обусловленности широко используется для матриц с линейными элементарными делителями.

Спектральное число обусловленности матрицы по отношению к проблеме собственных значений

30. Пусть A — матрица с линейными элементарными делителями и H — такая матрица, что

$$H^{-1}AH = \text{diag}(\lambda_i). \quad (30.1)$$

Если λ — собственное значение $(A + \varepsilon B)$, то матрица $(A + \varepsilon B - \lambda I)$ особенная, и, следовательно, ее определитель равен нулю. Тогда имеем

$$H^{-1}(A + \varepsilon B - \lambda I)H = \text{diag}(\lambda_i - \lambda) + \varepsilon H^{-1}BH \quad (30.2)$$

и, сосчитав определитель, видим, что матрица справа в (30.2) также должна быть особенной. Мы различаем два случая.

С л у ч а й 1. $\lambda = \lambda_i$ при некотором i .

С л у ч а й 2. $\lambda \neq \lambda_i$ при всех i , так что можно написать

$$\text{diag}(\lambda_i - \lambda) + \varepsilon H^{-1}BH = \text{diag}(\lambda_i - \lambda)[I + \varepsilon \text{diag}(\lambda_i - \lambda)^{-1}H^{-1}BH], \quad (30.3)$$

и, сосчитав определители, снова видим, что матрица в скобках должна быть особенной. Но если $(I + x)$ — особенная, мы должны иметь $\|x\| \geq 1$ для всех норм, так как если $\|x\| < 1$, ни одно из собственных значений $(I + x)$ не может быть нулем. Следовательно, мы имеем, в частности,

$$\|\varepsilon \text{diag}(\lambda_i - \lambda)^{-1}H^{-1}BH\|_2 \geq 1, \quad (30.4)$$

что дает

$$\varepsilon \max |(\lambda_i - \lambda)^{-1}| \|H^{-1}\|_2 \|B\|_2 \|H\|_2 \geq 1, \quad (30.5)$$

т. е.

$$\min |\lambda_i - \lambda| \leq \varepsilon \|H^{-1}\|_2 \|H\|_2 \|B\|_2. \quad (30.6)$$

Следовательно, в любом случае

$$|\lambda_i - \lambda| \leq \varepsilon k(H) \|B\|_2 \quad (30.7)$$

по крайней мере при одном значении i , где

$$k(H) = \|H^{-1}\|_2 \|H\|_2. \quad (30.8)$$

Ввиду зависимости $k(H)$ от спектральной нормы матрицы H его часто называют *спектральным числом обусловленности* (см. гл. 4, § 3). Мы показали, что общая чувствительность собственных значений A зависит от величины $k(H)$, так что $k(H)$ можно рассматривать как число обусловленности A по отношению к проблеме собственных значений. Результат, который мы доказали, принадлежит Бауэру и Файку (1960).

Заметим, что (30.6) справедливо для любой нормы, для которой

$$\|\text{diag}(\lambda_i - \lambda)^{-1}\| = \max |\lambda_i - \lambda|^{-1}, \quad (30.9)$$

и, следовательно, это верно как для $\|\cdot\|_1$, так и для $\|\cdot\|_\infty$. Так как это верно для $\|\cdot\|_2$, это верно *тем самым* для евклидовой нормы.

Используя понятие непрерывности так же, как при доказательстве второй теоремы Гершгорина, мы можем более аккуратно локализовать корни. Это приведет к следующему результату. Если s кругов

$$|\lambda_i - \lambda| \leq \varepsilon k(H) \|B\|_2 \quad (30.10)$$

образуют связную область, которая изолирована от остальных, то в этой области находится точно s собственных значений. *Заметим, что в этом результате мы не требуем, чтобы ε было мало.*

Свойства спектрального числа обусловленности

31. Так как матрица H не единственна (даже если собственные значения различны, каждый столбец может быть умножен на произвольный множитель), будем считать, что спектральное число обусловленности A по отношению к проблеме собственных значений — это наименьшее

значение $k(H)$ для всех допустимых H . В любом случае имеем

$$k(H) = \|H^{-1}\|_2 \|H\|_2 \geq \|H^{-1}H\|_2 = 1. \quad (31.1)$$

Если A — нормальная (гл. 1, § 48) и, в частности, если она эрмитова или унитарная, мы можем взять H унитарной, и тогда

$$k(H) = 1. \quad (31.2)$$

Проблема собственных значений всегда поэтому хорошо обусловлена для нормальных матриц, хотя это не обязательно справедливо для проблемы собственных векторов.

Рассмотрим теперь соотношение между $k(H)$ и величинами $|s_i^{-1}|$, которые управляют чувствительностью отдельных собственных значений. Нормированные правые и левые собственные векторы, соответствующие λ_i , имеют вид

$$x_i = \frac{He_i}{\|He_i\|_2}, \quad y_i = \frac{(H^{-1})^T e_i}{\|(H^{-1})^T e_i\|_2}. \quad (31.3)$$

Следовательно,

$$|s_i| = |y_i^T x_i| = \frac{|e_i^T H^{-1} H e_i|}{\|He_i\|_2 \|(H^{-1})^T e_i\|_2} = \frac{1}{\|He_i\|_2 \|(H^{-1})^T e_i\|_2}, \quad (31.4)$$

и мы имеем

$$\|He_i\|_2 \leq \|H\|_2 \|e_i\|_2 = \|H\|_2, \quad (31.5)$$

$$\|(H^{-1})^T e_i\|_2 \leq \|(H^{-1})^T\|_2 \|e_i\|_2 = \|(H^{-1})^T\|_2 = \|H^{-1}\|_2, \quad (31.6)$$

что дает

$$|s_i^{-1}| \leq k(H). \quad (31.7)$$

С другой стороны, мы можем взять столбцы H в виде $x_i/s_i^{1/2}$ и строки H^{-1} в виде $y_i^T/s_i^{1/2}$, и при таком выборе H

$$\begin{aligned} k(H) &= \|H\|_2 \|H^{-1}\|_2 \leq \|H\|_E \|H^{-1}\|_E = \\ &= \left(\sum_1^n |s_r^{-1}|\right)^{1/2} \left(\sum_1^n |s_r^{-1}|\right)^{1/2} = \sum_1^n |s_r^{-1}|. \end{aligned} \quad (31.8)$$

Для полного знания чувствительности собственных значений A по отношению к возмущениям ее элементов нам требуется n^3 величин $\frac{\partial \lambda_i}{\partial a_{kj}}$, так как чувствительность отдельных собственных значений по отношению к возмущениям отдельных элементов A может меняться в широких пределах. Такой большой агрегат невозможно применять в практической работе. Подходящим компромиссом будет введение n чисел $|s_i^{-1}|$, и мы будем называть их n числами обусловленности матрицы A по отношению к проблеме собственных значений. Надо, однако, заметить, что для того, чтобы найти приближенное значение $k(H)$, на практике мы обычно должны вычислить систему приближенных собственных векторов. Получив их, легче получить приближения к отдельным s_i , чем приближение к $k(H)$.

Инвариантные свойства чисел обусловленности

32. Как k , так и n чисел обусловленности $|s_i^{-1}|$ обладают важными свойствами инвариантности по отношению к унитарным преобразованиям подобия. Действительно, если R унитарная и

$$B = R A R^H, \quad (32.1)$$

то соответственно для

$$H^{-1}AH = \text{diag}(\lambda_i) \quad (32.2)$$

имеем

$$H^{-1}R^HBRH = \text{diag}(\lambda_i) \quad \text{или} \quad (RH)^{-1}B(RH) = \text{diag}(\lambda_i). \quad (32.3)$$

Следовательно, спектральное число обусловленности k' матрицы B равно

$$k' = \|(RH)^{-1}\|_2 \|RH\|_2 = \|H^{-1}\|_2 \|H\|_2 = k. \quad (32.4)$$

Аналогично, если x_i и y_i — правые и левые собственные векторы A , то собственные векторы B имеют вид

$$x'_i = Rx_i, \quad y'_i = \bar{R}y_i,$$

и, следовательно,

$$s'_i = (y'_i)^T x'_i = y_i^T \bar{R}^T R x_i = y_i^T R^H R x_i = y_i^T x_i = s_i. \quad (32.5)$$

Таким образом, чувствительность отдельных собственных значений инвариантна при унитарных преобразованиях.

Очень плохо обусловленные матрицы

33. Собственные значения, соответствующие нелинейным элементарным делителям, нужно рассматривать, вообще говоря, как плохо обусловленные, хотя следует помнить рассмотрение § 26. Однако мы не должны думать, что это главная форма плохой обусловленности. Даже если собственные значения различны и хорошо отделены друг от друга, они все же могут быть очень плохо обусловлены. Это хорошо иллюстрируется следующим примером.

Рассмотрим матрицу A 20-го порядка вида

$$A = \begin{bmatrix} 20 & 20 & & & & \\ & 19 & 20 & & & \\ & & 18 & 20 & & \\ & & & \dots & \dots & \\ & & & & 2 & 20 \\ & & & & & 1 \end{bmatrix}. \quad (33.1)$$

Это треугольная матрица, и поэтому ее собственными значениями являются ее диагональные элементы. Если добавить элемент ε к элементу в позиции (20,1), то характеристическое уравнение будет

$$(20 - \lambda)(19 - \lambda) \dots (1 - \lambda) = 20^{19}\varepsilon. \quad (33.2)$$

Мы видели в § 9, что при достаточно малых ε возмущение собственного значения $\lambda = r$ не содержит дробных степеней ε , и если мы напишем

$$\lambda_r(\varepsilon) - r \sim K_r \varepsilon, \quad (33.3)$$

то из (33.2) очевидно, что

$$K_r = \frac{20^{19}(-1)^r}{(20-r)!(r-1)!}. \quad (33.4)$$

Эта постоянная велика при каждом значении r . Наименьшие по модулю величины — K_1 и K_{20} , а наибольшие — K_{10} и K_{11} . На самом деле,

$$-K_1 = K_{20} \approx 10^7(4,31), \quad -K_{11} = K_{10} \approx 10^{12}(3,98). \quad (33.5)$$

Величины K_i так велики, что линеаризованная теория приложима лишь при очень малых ε . Для $\varepsilon = 10^{-10}$ собственные значения помещены в табл. 1, причем они размещены так, что видна симметрия вокруг значения 10,5.

Таблица 1

Собственные значения возмущенной матрицы ($\varepsilon=10^{-10}$)

0,99575439	3,96533070 $\pm i1,08773570$
20,00424561	17,03466930 $\pm i1,08773570$
2,10924184	5,89397755 $\pm i1,94852927$
18,89075816	15,10602245 $\pm i1,94852927$
2,57488140	8,11807338 $\pm i2,52918173$
18,42511860	12,88192662 $\pm i2,52918173$
	10,50000000 $\pm i2,73339736$

34. Изменения собственных значений, соответствующие этому единственному возмущению, так же велики, как и изменения собственных значений $C_{10}(a)$ при аналогичном возмущении, хотя последняя матрица имеет один элементарный делитель десятого порядка. Тот факт, что возмущения в конечном счете имеют порядок ε для первой и порядок $\varepsilon^{1/10}$ для второй при $\varepsilon \rightarrow 0$, не становится существенным до тех пор, пока мы не рассматриваем возмущения, значительно меньшие, чем 10^{-10} .

Малая устойчивость собственных значений матрицы (33.1) показывает, что s_i должны быть малыми. На самом деле, собственный вектор x_r матрицы A , соответствующий $\lambda = r$, имеет компоненты

$$\left[1, \frac{20-r}{-20}, \frac{(20-r)(19-r)}{(-20)^2}, \dots, \frac{(20-r)!}{(-20)^{20-r}}, 0, \dots, 0 \right], \quad (34.1)$$

а компоненты y_r равны

$$\left[0, \dots, 0, \frac{(r-1)!}{20^{r-1}}, \dots, \frac{(r-1)(r-2)}{20^2}, \frac{r-1}{20}, 1 \right]. \quad (34.2)$$

Эти векторы не нормированы по 2-норме, но $y_r^T x_r$ дает хорошую оценку величины s_r . Мы имеем на самом деле

$$|y_r^T x_r| = \frac{(20-r)!(r-1)!}{20^{19}}, \quad (34.3)$$

и обратная величина от правой части (34.3) в точности есть $|K_r|$ из уравнения (33.4). Заметим, что соответствующие собственные векторы A и A^T почти ортогональны. Хотя строгая ортогональность возможна лишь при нелинейных делителях, мы можем, очевидно, подойти очень близко к ней, даже когда матрица имеет хорошо отделенные собственные значения.

Можно ожидать, что существует матрица, близкая к A и имеющая нелинейные элементарные делители. Легко можно проверить, что это верно. В самом деле, мы можем построить такую матрицу возмущением элемента в позиции (20.1). Соответствующие возмущению ε собственные значения являются корнями уравнения (33.2). Функция в левой части этого уравнения симметрична относительно $\lambda = 10,5$ и имеет 9 максимумов и 10 минимумов. Если мы увеличиваем ε , начиная с нуля, то корни 10 и 11 двигаются навстречу друг другу и обязательно совпадут на 10,5 при

$$\left(\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{19}{2} \right)^2 = 20^{19} \varepsilon, \quad (34.4)$$

что дает для ε приблизительно $7,8 \cdot 10^{-14}$. Соответствующий элементарный делитель должен быть квадратичным, так как возмущенная матрица $(A - \lambda I)$ имеет ранг 19 при любых λ , так как матрица 19-го порядка в верхнем правом углу имеет минор, равный 20^{19} . Если ε увеличивать дальше, то корни 8 и 9 и 12 и 13 начнут двигаться навстречу друг другу, и из симметрии они совпадут при некотором значении ε . Матрица при этом имеет два двойных собственных значения, и каждому соответствует квадратичный элементарный делитель. Дальнейшее увеличение ε ведет к совпадению корней 6 и 7, 14 и 15 соответственно и т. д.

35. Все собственные значения матрицы, которую мы рассмотрели, были плохо обусловлены, хотя некоторые были хуже обусловлены, чем другие. Сейчас мы дадим пример множества матриц, каждая из которых имеет собственные значения с широко меняющейся обусловленностью. Это матрица класса B_n вида

$$B_n = \begin{bmatrix} n & n-1 & n-2 & \dots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \dots & 3 & 2 & 1 \\ & n-2 & n-2 & \dots & 3 & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & 2 & 2 & 1 \\ & & & & & 1 & 1 \end{bmatrix}. \quad (35.1)$$

Наибольшее собственное значение этой матрицы очень хорошо обусловлено, а наименьшее очень плохо. При $n = 12$ первые несколько s_i порядка единицы, а последние три порядка 10^{-7} . При увеличении n наименьшие собственные значения становятся все более плохо обусловлены.

То, что некоторые из собственных значений должны быть очень чувствительны к малым изменениям некоторых матричных элементов, видно из следующего рассмотрения. Определитель матрицы равен единице при всех n , как можно показать простыми операциями со строками матрицы. Если элемент в позиции $(1, n)$ заменен на $(1 + \varepsilon)$, то определитель будет

$$1 - (n-1)! \varepsilon. \quad (35.2)$$

Если $\varepsilon = 10^{-10}$ и $n = 20$, определитель изменится от 1 до $(1 - 19! \cdot 10^{-10})$, что равно приблизительно $-1,216 \cdot 10^7$. Но определитель матрицы равен произведению ее собственных значений. По крайней мере одно собственное значение возмущенной матрицы должно быть сильно отлично от собственного значения исходной. При $n = 12$ возмущение, меньшее 10^{-9} элемента в позиции $(1, 12)$, приводит к появлению квадратичного делителя.

Плохая обусловленность того типа, который мы сейчас обсудили, на наш взгляд гораздо более важна, чем плохая обусловленность, связанная с нелинейными элементарными делителями. Матриц, имеющих в точности нелинейные элементарные делители, почти не существует на практике. Даже в теоретических работах это главным образом матрицы специального вида, обычно с небольшими целыми коэффициентами. Если элементы таких матриц иррациональны или не могут быть представлены на применяемой цифровой вычислительной машине точно, ошибки округления обычно приводят к матрице, уже не имеющей нелинейных элементарных делителей. С другой стороны, матрицы с некоторыми малыми s_i весьма часто встречаются.

Теория возмущений для вещественных симметричных матриц

36. Несколько следующих параграфов посвящены развитию теории возмущений для вещественных симметричных матриц. Эта теория проще общей теории, так как две системы векторов x_i и y_i можно взять одинаковыми (даже в случае нескольких кратных собственных значений) и, следовательно, все s_i равны единице. Как мы уже заметили в § 31, проблема собственных значений для вещественных симметричных матриц всегда хорошо обусловлена. Некоторые из результатов, которые мы докажем в следующих параграфах, очевидным образом немедленно распространяются на весь класс нормальных матриц, некоторые распространяются лишь на эрмитовы матрицы или не распространяются вообще. Мы будем без доказательств указывать возможные распространения в каждом случае.

Несимметричные возмущения

37. Иногда нас будут интересовать несимметричные возмущения симметричной матрицы, и соответственно мы изучим систему собственных значений $(A + \varepsilon B)$, где A симметрична, а B нет. Из (30.6) мы знаем, что каждое собственное значение $(A + \varepsilon B)$ лежит по крайней мере в одном из кругов

$$|\lambda_i - \lambda| \leq \varepsilon \|H^{-1}\|_2 \|H\|_2 \|B\|_2, \quad (37.1)$$

и так как H можно взять ортогональной, то

$$|\lambda_i - \lambda| \leq \varepsilon \|B\|_2 \leq n\varepsilon \quad (\text{если } |b_{ij}| \leq 1). \quad (37.2)$$

Мы знаем, далее, что если s из этих кругов образуют связную область, изолированную от остальных, то в этой области находятся ровно s собственных значений. Эти результаты, очевидно, справедливы для всех нормальных матриц.

В случаях вещественных симметричных матриц матрица B обычно вещественна, и, следовательно, каждое комплексное собственное значение $A + \varepsilon B$ должно встречаться в сопряженной паре. Если $|\lambda_i - \lambda_j| > 2n\varepsilon$ ($j \neq i$), то i -й круг изолирован и содержит лишь одно собственное значение. Это собственное значение поэтому должно быть вещественным. Следовательно, если все собственные значения A отделены более чем на $2n\varepsilon$, то случайное несимметричное возмущение в каждом элементе, по модулю меньшее ε , оставляет собственные значения вещественными и простыми, и, следовательно, собственные векторы также вещественны и образуют полную систему.

Симметричные возмущения

38. Большинство из наиболее точных способов вычисления собственных значений вещественных симметричных матриц основано на использовании ортогональных преобразований подобия. Точная симметрия сохраняется в последовательных преобразованных матрицах тем, что вычисляется только верхняя треугольная часть каждой матрицы. Элементы под диагональю полагают равными соответствующим элементам над диагональю. По этой причине мы более подробно рассмотрим симметричные возмущения симметричных матриц.

Если возмущение симметрично, возмущенная матрица обязательно имеет вещественную систему собственных значений, и это же верно для возмущающей матрицы. Естественно искать соотношение между собственными значениями исходной, возмущающей и возмущенной матриц. Большинство результатов, полученных ранее, требовало малости возмущения. Увидим, с другой стороны, что многие результаты, которые мы докажем, свободны от этого ограничения, и потому, опуская ϵ , запишем

$$C = A + B, \quad (38.1)$$

где A , B и C вещественны и симметричны.

Классическая техника

39. Для исследования собственных значений суммы двух симметричных матриц существует весьма эффективная техника, основанная на принципе минимакса. Частично с целью продемонстрировать ее силу, мы начинаем с анализа, основанного на более классических методах.

Сначала выведем простой результат для окаймленных диагональных матриц. Пусть симметричная матрица X имеет вид

$$X = \left[\begin{array}{c|c} \alpha & a^T \\ \hline a & \text{diag}(\alpha_i) \end{array} \right] \quad (i = 1, \dots, n-1), \quad (39.1)$$

где a — вектор $(n-1)$ -й размерности. Мы хотим найти соотношения между собственными значениями X и числами α_i .

Предположим, что лишь s компонент a не нули; если a_j — нуль, то α_j — собственное значение X . Подходящим выбором матрицы перестановки P , затрагивающей лишь последние $(n-1)$ строк, можем получить матрицу Y такую, что

$$Y = P^T X P = \left[\begin{array}{c|c|c} \alpha & b^T & 0 \\ \hline b & \text{diag}(\beta_i) & 0 \\ \hline 0 & 0 & \text{diag}(\gamma_i) \end{array} \right], \quad (39.2)$$

где компоненты b не равны нулю, $\text{diag}(\beta_i)$ порядка s , $\text{diag}(\gamma_i)$ порядка $n-s-1$ и β_i и γ_i вместе составляют перестановку α_i . Заметим, что некоторые γ_i могут быть кратными собственными значениями $\text{diag}(\alpha_i)$.

Поэтому собственными значениями X будут γ_i , а также собственные значения матрицы Z вида

$$Z = \left[\begin{array}{c|c} \alpha & b^T \\ \hline b & \text{diag}(\beta_i) \end{array} \right]. \quad (39.3)$$

Если $s = 0$, то Z состоит из единственного элемента α и, следовательно, собственные значения X совпадают с собственными значениями $\text{diag}(\alpha_i)$ и величиной α . В противном случае рассмотрим характеристический полином Z , который равен

$$(\alpha - \lambda) \prod_{i=1}^s (\beta_i - \lambda) + \sum_{j=1}^s b_j^2 \prod_{i \neq j} (\beta_i - \lambda) = 0. \quad (39.4)$$

Предположим, что только t значений из β_i различны, и обозначим их $\beta_1, \beta_2, \dots, \beta_t$, а их кратности $r_1, r_2, \dots, r_t$ соответственно, так что

$$r_1 + r_2 + \dots + r_t = s. \quad (39.5)$$

(Может случиться, конечно, что $t = s$ и все β_i простые.) Тогда ясно, что левая часть (39.4) имеет множитель

$$\prod_{i=1}^t (\beta_i - \lambda)^{r_i - 1}, \quad (39.6)$$

так что β_i является собственным значением Z кратности $r_i - 1$.

Поделив (39.4) на $\prod_{i=1}^t (\beta_i - \lambda)^{r_i}$, видим, что остальные собственные значения Z являются корнями уравнения

$$0 = \alpha - \lambda - \sum_{i=1}^t c_i^2 (\beta_i - \lambda)^{-1} \equiv \alpha - f(\lambda), \quad (39.7)$$

где каждый коэффициент c_i^2 — это сумма r_i величин b_{ij}^2 , связанных с β_i , и поэтому строго положителен. График $f(\lambda)$ дан на рис. 1, где мы расположили различные β_i в убывающем порядке.

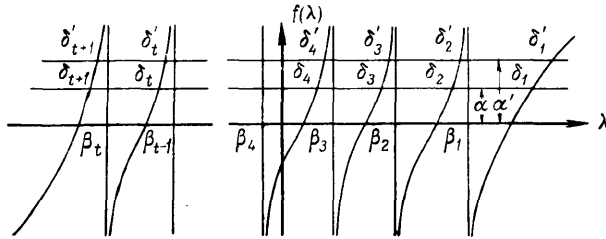


Рис. 1.

Очевидно, что $(t + 1)$ корней уравнения $\alpha = f(\lambda)$, которые мы обозначим $\delta_1, \delta_2, \dots, \delta_{t+1}$, удовлетворяют соотношениям

$$\left. \begin{aligned} \infty > \delta_1 > \beta_1; \beta_{i-1} > \delta_i > \beta_i \quad (i = 2, 3, \dots, t); \\ \beta_t > \delta_{t+1} > -\infty. \end{aligned} \right\} \quad (39.8)$$

Поэтому n собственных значений X распадутся на три множества:

(i) Собственные значения $\gamma_1, \gamma_2, \dots, \gamma_{n-1-s}$, соответствующие нулевым α_i . Они равны $(n - 1 - s)$ числам из α_i .

(ii) $(s - t)$ собственных значений, состоящих из $(r_i - 1)$ величин, равных β_i ($i = 1, 2, \dots, t$). Они равны еще $(s - t)$ числам из α_i .

(iii) $(t + 1)$ собственных значений, равных δ_i , удовлетворяющих соотношению (39.8). Если $t = 0$, то $\delta_1 = \alpha$.

Заметим, что одно или оба множества (i) и (ii) могут быть пустыми. В любом случае элементы этих множеств не зависят от α . Если собственные значения X обозначить λ_i ($i = 1, 2, \dots, n$) и расположить в невозрастающем порядке, то, если α_i также расположены в невозрастающем порядке, немедленно следует, что

$$\lambda_1 \geq \alpha_1 \geq \lambda_2 \geq \alpha_2 \geq \dots \geq \alpha_{n-1} \geq \lambda_n. \quad (39.9)$$

Другими словами, α_i разделяют λ_i по крайней мере в слабом смысле.

40. Рассмотрим теперь собственные значения матрицы X' , полученной из X заменой α на α' . Собственные значения из множеств (i) и (ii) для X и X' совпадают. Обозначим собственные значения X' из множества (iii)

$\delta'_1, \delta'_2, \dots, \delta'_{t+1}$. Далее, для положительных λ имеем

$$\frac{df}{d\lambda} = 1 + \sum_1^t \frac{c_i^2}{(\beta_i - \lambda)^2} > 1 \quad \text{при всех } \lambda, \quad (40.1)$$

и, следовательно, каждая разность $\delta'_i - \delta_i$ лежит между 0 и $\alpha' - \alpha$ (см. рис. 1). Мы можем поэтому положить

$$\delta'_i - \delta_i = m_i (\alpha' - \alpha), \quad (40.2)$$

где

$$0 < m_i < 1 \quad \text{и} \quad \sum m_i = 1. \quad (40.3)$$

Если $t = 0$, то

$$\delta'_1 = \alpha', \quad \delta_1 = \alpha, \quad \text{и} \quad \delta'_1 - \delta_1 = \alpha' - \alpha. \quad (40.4)$$

Следовательно, можем написать во всех случаях

$$\delta'_i - \delta_i = m_i (\alpha' - \alpha), \quad (40.5)$$

где

$$0 < m_i \leq 1, \quad \sum m_i = 1. \quad (40.6)$$

Так как другие собственные значения X и X' равны, можно сказать, что мы установили соответствие между n собственными значениями X и X' , которые переименовали в $\lambda_1, \lambda_2, \dots, \lambda_n$ и $\lambda'_1, \lambda'_2, \dots, \lambda'_n$, так что

$$\left. \begin{aligned} \lambda'_i - \lambda_i &= m_i (\alpha' - \alpha), \\ 0 &\leq m_i \leq 1, \quad \sum m_i = 1, \end{aligned} \right\} \quad (40.7)$$

причем, очевидно, $m_i = 0$ для собственных значений из множеств (i) и (ii).

Более того, если соотношения (40.7) выполняются при некотором порядке расположения λ_i и λ'_i , то они *тем более* выполняются, если как λ_i , так и λ'_i расположены в невозрастающем порядке.

Симметричная матрица с рангом единица

41. Результаты предыдущего параграфа могут быть применены для оценки собственных значений матрицы

$$C = A + B, \quad (41.1)$$

где A и B симметричные и ранг B равен единице. В этом случае существует ортогональная матрица R такая, что

$$R^T B R = \begin{bmatrix} \rho & & 0 \\ & \ddots & \\ 0 & & 0 \end{bmatrix}, \quad (41.2)$$

где ρ — единственное отличное от нуля собственное значение матрицы B . Если мы обозначим

$$R^T A R = \begin{bmatrix} \alpha & & a^T \\ & \ddots & \\ a & & A_{n-1} \end{bmatrix}, \quad (41.3)$$

то существует ортогональная матрица S порядка $(n-1)$ такая, что

$$S^T A_{n-1} S = \text{diag}(\alpha_i). \quad (41.4)$$

Если мы определим далее Q соотношением

$$Q = R \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & S \end{array} \right], \quad (41.5)$$

то Q ортогональна и

$$Q^T (A + B) Q = \left[\begin{array}{c|c} \alpha & b^T \\ \hline b & \text{diag}(\alpha_i) \end{array} \right] + \left[\begin{array}{c|c} \rho & 0 \\ \hline 0 & 0 \end{array} \right], \quad (41.6)$$

где $b = S^T a$. Собственные значения A и $A + B$ поэтому совпадают с собственными значениями

$$\left[\begin{array}{c|c} \alpha & b^T \\ \hline b & \text{diag}(\alpha_i) \end{array} \right] \quad \text{и} \quad \left[\begin{array}{c|c} \alpha + \rho & b^T \\ \hline b & \text{diag}(\alpha_i) \end{array} \right], \quad (41.7)$$

и, если мы расположим эти собственные значения в убывающем порядке, то они будут удовлетворять соотношениям

$$\left. \begin{array}{l} \lambda'_i - \lambda = m_i \rho \\ 0 \leq m_i \leq 1, \quad \sum m_i = 1. \end{array} \right\} \quad \text{при} \quad (41.8)$$

Следовательно, если мы добавим B к A , то все собственные значения A изменятся на величины, лежащие между нулем и собственным значением ρ матрицы B .

Кроме того, из замечания в конце § 39 и из преобразования, данного в (41.7), очевидно, следует, что собственные значения главного минора A_{n-1} матрицы A разделяют собственные значения A по крайней мере в слабом смысле. Если A имеет собственное значение λ_i кратности k , то A_{n-1} должна иметь собственное значение λ_i кратности $k - 1$, k или $k + 1$.

Экстремальные свойства собственных значений

42. Прежде чем дать минимаксные характеристики собственных значений, рассмотрим проблему определения максимального значения M функции $x^T A x$ при ограничениях $x^T x = 1$. Ясно, что M также удовлетворяет соотношению

$$M = \max_{x \neq 0} \left(\frac{x^T A x}{x^T x} \right). \quad (42.1)$$

Так как A — симметричная матрица, существует ортогональная матрица R такая, что

$$R^T A R = \text{diag}(\lambda_i). \quad (42.2)$$

Введем новые переменные y , полагая

$$x = R y, \quad \text{т. е.} \quad y = R^T x. \quad (42.3)$$

Тогда

$$x^T A x = y^T R^T A R y = y^T \text{diag}(\lambda_i) y = \sum_{i=1}^n \lambda_i y_i^2, \quad (42.4)$$

$$x^T x = y^T R^T R y = \sum_{i=1}^n y_i^2. \quad (42.5)$$

Так как (42.3) устанавливает взаимно однозначное соответствие между x и y , начальная задача эквивалентна задаче нахождения максимального значения $\sum_1^n \lambda_i y_i^2$ при условии $\sum_1^n y_i^2 = 1$. Если мы предположим, что λ_i расположены в невозрастающем порядке, то ясно, что максимальное значение есть λ_1 , и оно достигается при $y = e_1$. Соответствующий вектор x равен r_1 , первому столбцу R , который является на самом деле собственным вектором A , соответствующим λ_1 . (Заметим, что если $\lambda_1 = \lambda_2 = \dots = \lambda_r \neq \lambda_{r+1}$, то любой вектор единичной длины из подпространства, натянутого на $e_1, e_2, \dots, e_r$, дает значение λ_1 .) Аналогично λ_n есть минимальное значение $x^T A x$ при том же условии.

Рассмотрим ту же проблему максимизации при дополнительном условии, что x ортогонален к r_1 . Из соотношения

$$0 = r_1^T x = r_1^T R y = e_1^T R^T R y = e_1^T y \quad (42.6)$$

видим, что соответствующее ограничение для y означает, что он должен иметь нулевую первую компоненту. Следовательно, максимальным значением $x^T A x$ при таком ограничении будет λ_2 , и оно достигается при $y = e_2$ и, значит, при $x = r_2$. Аналогично мы находим, что максимальное значение $x^T A x$ при дополнительных ограничениях $r_1^T x = r_2^T x = \dots = r_s^T x = 0$ равно λ_{s+1} , и оно достигается при $x = r_{s+1}$.

Слабость такой характеристики собственных значений состоит в том, что определение каждого λ_s зависит от знания векторов $r_1, r_2, \dots, r_{s-1}$, которые являются собственными векторами A , соответствующими $\lambda_1, \lambda_2, \dots, \lambda_{s-1}$.

Минимаксные свойства собственных значений

43. Сейчас мы опишем характеристики собственных значений (см., например, Курант и Гильберт, 1953), свободные от недостатка, о котором мы говорили. Рассмотрим максимальное значение $x^T A x$ при условиях

$$x^T x = 1, \quad p_i^T x = 0, \quad p_i \neq 0 \quad (i = 1, 2, \dots, s; \quad s < n), \quad (43.1)$$

где p_i — произвольные ненулевые векторы. Другими словами, рассмотрим лишь значения x , удовлетворяющие s линейным условиям. Для всех значений x имеем

$$\lambda_n \leq x^T A x \leq \lambda_1, \quad (43.2)$$

так что $x^T A x$ ограничена, и ее максимум является функцией n s компонент p_i . Теперь зададим вопрос: какое минимальное значение возможно для этого максимума при всевозможных выборах s векторов p_i ?

Как и раньше, мы можем работать с y , определенными в (42.3). Соотношения (43.1) станут

$$y^T y = 1, \quad q_i^T y = 0, \quad q_i^T = p_i^T R \neq 0. \quad (43.3)$$

Рассмотрим некоторый частный выбор $p_1, p_2, \dots, p_s$. Он даст нам соответствующую систему q_i и s линейных однородных уравнений для n переменных y_j . Если мы добавим соотношения

$$y_{s+2} = y_{s+3} = \dots = y_n = 0, \quad (43.4)$$

то получим $(n - 1)$ однородное уравнение с n переменными $y_1, y_2, \dots, y_n$, и, следовательно, существует по крайней мере одно ненулевое решение

$$(y_1, y_2, \dots, y_{s+1}, 0, \dots, 0), \quad (43.5)$$

которое можно нормировать так, что $\sum_{i=1}^{s+1} y_i^2 = 1$. При таком выборе y имеем

$$y^T \text{diag}(\lambda_i) y = \sum_{i=1}^{s+1} \lambda_i y_i^2 \geq \lambda_{s+1}. \quad (43.6)$$

Это показывает, что при любом выборе p_i существует y , и поэтому x , такой, для которого

$$x^T A x = y^T \text{diag}(\lambda_i) y \geq \lambda_{s+1}, \quad (43.7)$$

и, следовательно,

$$\max (x^T A x) \geq \lambda_{s+1}. \quad (43.8)$$

Это означает, что

$$\min \max x^T A x \geq \lambda_{s+1}. \quad (43.9)$$

Однако если мы возьмем $p_i = Re_i$, то $q_i^T = e_i^T$ и соотношения (43.3) станут

$$y_i = 0 \quad (i = 1, 2, \dots, s), \quad (43.10)$$

и, следовательно, для любых y при таком частном выборе имеем

$$x^T A x = y^T \text{diag}(\lambda_i) y = \sum_{i=1}^n \lambda_i y_i^2 \leq \lambda_{s+1}, \quad (43.11)$$

что дает

$$\max (x^T A x) \leq \lambda_{s+1}. \quad (43.12)$$

Соотношения (43.8) и (43.12) вместе означают, что при таком выборе p_i

$$\max (x^T A x) = \lambda_{s+1}, \quad (43.13)$$

и это значение достигается при y вида

$$y_{s+1} = 1, \quad y_i = 0 \quad (i \neq s+1). \quad (43.14)$$

Следовательно, $\min \max (x^T A x)$, когда x удовлетворяет s линейным соотношениям, равен λ_{s+1} . Этот результат известен как теорема Куранта—Фишера. Мы подчеркиваем, что показано существование по крайней мере одной системы p_i и одного соответствующего x , для которых это минимальное значение достигается.

Совершенно аналогичным образом можем доказать, что

$$\lambda_s = \max \min (x^T A x) \quad (43.15)$$

при

$$x^T x = 1, \quad p_i^T x = 0 \quad (i = 1, 2, \dots, n-s).$$

Заметим, что в обеих характеристиках среди множеств p_i есть такие, в которых некоторые p_i равны, хотя, вообще говоря, действительная минимизация максимума достигается для системы различных p_i . Если мы имеем любые s векторов p_i , то можно утверждать, что

$$\lambda_{s+1} \leq \max (x^T A x) \quad (43.16)$$

при

$$x^T x = 1, \quad p_i^T x = 0 \quad (i = 1, 2, \dots, s),$$

даже если не все p_i различны.

Собственные значения суммы двух симметричных матриц

44. Минимаксную характеристику собственных значений можно использовать для установления соотношения между собственными значениями симметричных матриц A , B и C таких, что

$$C = A + B. \quad (44.1)$$

Обозначим собственные значения A , B и C через α_i , β_i и γ_i соответственно, располагая все три множества в невозрастающем порядке. По теореме минимакса имеем

$$\left. \begin{aligned} \gamma_s &= \min \max (x^T C x), \\ x^T x &= 1, \quad p_i^T x = 0 \quad (i = 1, 2, \dots, s-1). \end{aligned} \right\} \quad (44.2)$$

Следовательно, при некотором выборе p_i для всех соответствующих x имеем

$$\gamma_s \leq \max (x^T C x) = \max (x^T A x + x^T B x). \quad (44.3)$$

Если R — ортогональная матрица такая, что

$$R^T A R = \text{diag}(\alpha_i), \quad (44.4)$$

и если мы возьмем $p_i = R e_i$, то

$$0 = p_i^T x = e_i^T y \quad (i = 1, 2, \dots, s-1). \quad (44.5)$$

При таком выборе p_i первые $(s-1)$ компонент y равны нулю, и из (44.3) имеем

$$\gamma_s \leq \max (x^T A x + x^T B x) = \max \left(\sum_{i=s}^n \alpha_i y_i^2 + x^T B x \right). \quad (44.6)$$

Однако

$$\sum_{i=s}^n \alpha_i y_i^2 \leq \alpha_s, \quad (44.7)$$

в то время как

$$x^T B x \leq \beta_1 \quad (44.8)$$

для всех x . Следовательно, выражение в скобках не больше чем $\alpha_s + \beta_1$ при всех x , соответствующих этому выбору p_i . Поэтому, его максимум не больше чем $\alpha_s + \beta_1$, и мы имеем

$$\gamma_s \leq \alpha_s + \beta_1. \quad (44.9)$$

Так как $A = C + (-B)$ и собственные значения $(-B)$ в невозрастающем порядке равны $-\beta_n, -\beta_{n-1}, \dots, -\beta_1$, приложение только что доказанного результата дает

$$\alpha_s \leq \gamma_s + (-\beta_n) \quad \text{или} \quad \gamma_s \geq \alpha_s + \beta_n. \quad (44.10)$$

Соотношения (44.9) и (44.10) показывают, что если B добавлена к A , все собственные значения меняются на величину, лежащую между наименьшим и наибольшим собственными значениями B . Это находится в согласии с результатом, доказанным в § 41 аналитическими методами для матриц B ранга единица. Заметим, что мы не предполагали здесь малости возмущения и что результаты не зависят от кратностей собственных значений A , B и C .

Практические применения

45. Результат предыдущего параграфа один из наиболее используемых на практике. Часто B будет мала, и единственная информация, которую мы будем иметь, это некоторая верхняя граница для ее элементов или для некоторых норм. Если, например, мы имеем

$$|b_{ij}| \leq \varepsilon, \quad (45.1)$$

то

$$-n\varepsilon \leq \beta_n \leq \beta_1 \leq n\varepsilon, \quad (45.2)$$

и, следовательно,

$$|\gamma_r - \alpha_r| \leq n\varepsilon. \quad (45.3)$$

Это более строго, чем наш ранний результат § 37 для несимметричных возмущений, так как теперь нет ограничения в разделении α_i , β_i и γ_i . Раньше мы могли доказать соотношение (45.3) только в том случае, когда собственные значения α_r все отделены более чем на $2n\varepsilon$.

Дальнейшие применения принципа минимакса

46. Результат § 44 может быть обобщен в виде

$$\gamma_{r+s-1} \leq \alpha_r + \beta_s \quad (r+s-1 \leq n). \quad (46.1)$$

Доказательство. Существует по крайней мере одна система p_i , для которых

$$\max(x^T A x) = \alpha_r \quad \text{при условии} \quad p_i^T x = 0 \quad (i=1, 2, \dots, r-1), \quad (46.2)$$

и система q_i , для которых

$$\max(x^T B x) = \beta_s \quad \text{при условии} \quad q_i^T x = 0 \quad (i=1, 2, \dots, s-1). \quad (46.3)$$

Рассмотрим множество x , удовлетворяющих и $p_i^T x = 0$ и $q_i^T x = 0$. Такие ненулевые x существуют, так как общее число уравнений равно $r+s-2$, что по (46.1) не больше $n-1$. Для всех таких x имеем

$$x^T C x = x^T A x + x^T B x \leq \alpha_r + \beta_s. \quad (46.4)$$

Следовательно,

$$\max(x^T C x) \leq \alpha_r + \beta_s \quad (46.5)$$

для всех x , удовлетворяющих некоторой системе $r+s-2$ линейных уравнений (46.2) и (46.3). Поэтому

$$\gamma_{r+s-1} = \min \max(x^T C x) \leq \alpha_r + \beta_s, \quad (46.6)$$

где минимизация проведена по всем множествам $r+s-2$ соотношений.

Теорема разделения

47. В качестве дальнейшей иллюстрации теоремы минимакса докажем, что собственные значения $\lambda'_1, \lambda'_2, \dots, \lambda'_{n-1}$ главного минора A_{n-1} матрицы A_n разделяют собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$ матрицы A . (Этот результат был доказан в § 41 аналитическими методами для нижнего главного минора порядка $(n-1)$; ясно, что это верно для всех

главных миноров порядка $(n - 1)$, что можно легко увидеть, применив преобразование подобия с матрицами перестановок.)

Множество значений $x^T A_{n-1} x$ для всех нормированных векторов $(n - 1)$ -й размерности совпадает с множеством значений $x^T A_n x$ для нормированных векторов n -й размерности при условии $x_n = 0$. Следовательно,

$$\left. \begin{aligned} \lambda'_s &= \min \max (x^T A_n x), \quad x^T x = 1, \\ x_n &= 0, \quad p_i^T x = 0 \quad (i = 1, 2, \dots, s - 1). \end{aligned} \right\} \quad (47.1)$$

Это максимальное значение будет достигаться при некотором выборе p_i . При этом выборе p_i λ'_s есть максимальное значение $x^T A_n x$ при s линейных соотношениях второй строки (47.1). Однако λ_{s+1} — это минимальное значение для максимума при любых s соотношениях. Следовательно,

$$\lambda_{s+1} \leq \lambda'_s. \quad (47.2)$$

Теперь рассмотрим любой выбор $(s - 1)$ векторов p_i . Обозначим максимальное значение $x^T A_n x$ для нормированных x , удовлетворяющих соответствующим линейным соотношениям, через $f_n(p_i)$. Пусть максимум при дополнительном условии $x_n = 0$ обозначен через $f_{n-1}(p_i)$. Тогда, очевидно,

$$f_{n-1}(p_i) \leq f_n(p_i) \quad (47.3)$$

и, следовательно,

$$\min f_{n-1}(p_i) \leq \min f_n(p_i), \quad (47.4)$$

что дает

$$\lambda'_s \leq \lambda_s. \quad (47.5)$$

Вся теория последних шести параграфов немедленно распространяется на эрмитовы матрицы; результаты и доказательства получаются заменой значка T на H всюду, где он встречается.

Теорема Виландта — Гофмана

48. Эта теорема несколько другого типа, чем те, которые мы рассматривали. Она связывает возмущения собственных значений с евклидовой нормой возмущающей матрицы. Теорема, принадлежащая Гофману и Виландту (1953), состоит в следующем.

Если $C = A + B$, где A , B и C — симметричные матрицы с собственными значениями α_i , β_i и γ_i , расположенными в невозрастающем порядке, то

$$\sum_{i=1}^n (\gamma_i - \alpha_i)^2 \leq \|B\|_E^2 = \sum_{i=1}^n \beta_i^2. \quad (48.1)$$

Гофман и Виландт дали весьма изящное доказательство, которое основано на теории линейного программирования. Доказательство, которое следует, менее изящно, но более элементарно. Оно в основном принадлежит Гивенсу (1954).

Нам понадобятся две простые леммы.

(i) Если симметричная матрица X имеет собственные значения λ_i , то

$$\|X\|_E^2 = \sum_{i=1}^n \lambda_i^2. \quad (48.2)$$

Действительно, существует ортогональная матрица R такая, что

$$R^T X R = \text{diag}(\lambda_i), \quad (48.3)$$

и, взяв евклидовы нормы от обеих частей, используя независимость евклидовой нормы от ортогональных преобразований, получим требуемый результат.

(ii) Если $R_1, R_2, \dots$ — бесконечная последовательность ортогональных матриц, то существует подпоследовательность $R_{s_1}, R_{s_2}, \dots$ такая, что

$$\lim_{i \rightarrow \infty} R_{s_i} = R. \quad (48.4)$$

Это следует из теоремы Больцано — Вейерштрасса для n^2 -мерного пространства, так как все элементы всех R_i лежат в промежутке $(-1, 1)$. Матрица R должна быть ортогональна, ибо

$$R_{s_i}^T R_{s_i} = I, \quad (48.5)$$

и потому, переходя к пределу, получаем

$$R^T R = I. \quad (48.6)$$

49. Пусть R_1 и R_2 — ортогональные матрицы такие, что

$$R_1^T B R_1 = \text{diag}(\beta_i), \quad R_2^T C R_2 = \text{diag}(\gamma_i). \quad (49.1)$$

Тогда

$$\begin{aligned} \text{diag}(\beta_i) &= R_1^T B R_1 = R_1^T (C - A) R_1 = \\ &= R_1^T [R_2 \text{diag}(\gamma_i) R_2^T - A] R_1 = R_1^T R_2 [\text{diag}(\gamma_i) - R_2^T A R_2] R_2^T R_1, \end{aligned} \quad (49.2)$$

и, взяв нормы от обеих сторон, получим

$$\sum_{i=1}^n \beta_i^2 = \|\text{diag}(\gamma_i) - R_2^T A R_2\|_E^2. \quad (49.3)$$

Рассмотрим теперь множество значений $\|\text{diag}(\gamma_i) - R^T A R\|_E^2 = f(R)$ для всех ортогональных матриц R . Уравнение (49.3) показывает, что

$\sum_{i=1}^n \beta_i^2$ принадлежит к этому множеству. Это множество, очевидно, ограничено и поэтому имеет конечные верхнюю и нижнюю границы u и l . Эти границы должны достигаться при некоторых R , так как $f(R)$ — непрерывная функция на компактном множестве ортогональных матриц.

50. Мы сейчас покажем, что l должна достигаться на таких R , для которых $R^T A R$ диагональна. Предположим, что существуют r различных γ_i , которые мы обозначим δ_i ($i = 1, 2, \dots, r$), расположенных так, что

$$\delta_1 > \delta_2 > \dots > \delta_r. \quad (50.1)$$

Мы можем написать

$$\text{diag}(\gamma_i) = \begin{bmatrix} \delta_1 I & & & \\ & \delta_2 I & & \\ & & \ddots & \\ & & & \delta_r I \end{bmatrix}, \quad (50.2)$$

где единичные матрицы имеют соответствующие порядки. Если $R^T A R$ разбить в соответствии с (50.2), то можно написать

$$R^T A R = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1r} \\ X_{21} & X_{22} & \dots & X_{2r} \\ \dots & \dots & \dots & \dots \\ X_{r1} & X_{r2} & \dots & X_{rr} \end{bmatrix} \equiv X. \quad (50.3)$$

Мы покажем сначала, что l может быть достигнута лишь при тех R , у которых недиагональные блоки в (50.3) — нули. Предположим, что существует ненулевой элемент x матрицы $R^T A R$, находящийся в строке p и столбце q , и что такой элемент находится в блоке X_{ij} ($i \neq j$). Тогда элементы на пересечении строк и столбцов p и q матриц $\text{diag}(\gamma_i)$ и $R^T A R$ имеют вид

$$\begin{array}{ccccccc} & \text{diag}(\gamma_i) & & & X = R^T A R & & \\ & \text{ст. } p & \text{ст. } q & & \text{ст. } p & \text{ст. } q & \\ \cdots & \delta_i & 0 & \cdots & a & x & \cdots \\ & \vdots & \vdots & & \vdots & \vdots & \\ \cdots & 0 & \delta_j & \cdots & x & b & \cdots \end{array} \quad (50.4)$$

стр. p стр. q стр. p стр. q

где для простоты мы опустили индексы у рассматриваемых элементов X . Мы покажем, что можно выбрать элементарную ортогональную матрицу S , соответствующую вращению в плоскости (p, q) , такую, что

$$q(S) = \|\text{diag}(\gamma_i) - S^T R^T A R S\|_E^2 - \|\text{diag}(\gamma_i) - R^T A R\|_E^2 < 0. \quad (50.5)$$

Обозначим a', x', b' соответствующие элементы $S^T R^T A R S$ на пересечении строк и столбцов p и q . Так как

$$\|S^T R^T A R S\|_E^2 = \|R^T A R\|_E^2, \quad (50.6)$$

легко видеть прямо из определения евклидовой нормы, что

$$q(S) = -2a'\delta_i - 2b'\delta_j + 2a\delta_i + 2b\delta_j = 2(a - a')\delta_i + 2(b - b')\delta_j, \quad (50.7)$$

и все остальные члены пропадают. Если угол вращения равен θ , то

$$\left. \begin{aligned} a' &= a \cos^2 \theta - 2x \cos \theta \sin \theta + b \sin^2 \theta, \\ b' &= a \sin^2 \theta + 2x \cos \theta \sin \theta + b \cos^2 \theta. \end{aligned} \right\} \quad (50.8)$$

Это дает

$$\begin{aligned} h(\theta) = q(S) &= 2\delta_i[(a - b) \sin^2 \theta + x \sin 2\theta] + \\ &+ 2\delta_j[(b - a) \sin^2 \theta - x \sin 2\theta] = P \sin^2 \theta + Q \sin 2\theta, \end{aligned} \quad (50.9)$$

где $Q \neq 0$, так как x не нуль по предположению и $\delta_i - \delta_j$ не нуль, так как δ_i и δ_j из различных диагональных блоков. Теперь имеем

$$\frac{d}{d\theta} h(\theta) = P \sin 2\theta + 2Q \cos 2\theta, \quad (50.10)$$

и, следовательно, производная равна $2Q$ при $\theta = 0$, что показывает, что подходящим выбором θ мы можем сделать $q(S) > 0$ и $q(S) < 0$.

51. Таким образом, мы показали, что $f(R)$ не может быть максимумом или минимумом при любом R , для которого X имеет ненулевые элементы

в каком-либо X_{ij} ($i \neq j$). Предположим теперь, что R — ортогональная матрица, для которой $f(R)$ достигает своего минимального значения. Тогда мы должны иметь

$$\text{diag}(\gamma_i) - R^T A R = \begin{bmatrix} \delta_1 I & & \\ & \delta_2 I & \\ & & \ddots \\ & & & \delta_r I \end{bmatrix} - \begin{bmatrix} X_{11} & & \\ & X_{22} & \\ & & \ddots \\ & & & X_{rr} \end{bmatrix}. \quad (51.4)$$

Если Q_i — ортогональные матрицы такие, что

$$Q_i^T X_{ii} Q_i = D_i, \quad (51.2)$$

где D_i — диагональные, то, обозначая прямую сумму Q_i через Q , получим $Q^T \text{diag}(\gamma_i) Q - Q^T R^T A R Q =$

$$= \begin{bmatrix} \delta_1 I & & \\ & \delta_2 I & \\ & & \ddots \\ & & & \delta_r I \end{bmatrix} - \begin{bmatrix} D_1 & & \\ & D_2 & \\ & & \ddots \\ & & & D_r \end{bmatrix} = \\ = \text{diag}(\gamma_i) - Q^T R^T A R Q. \quad (51.3)$$

Следовательно,

$$f(RQ) = f(R), \quad (51.4)$$

и минимум должен всегда достигаться на матрице RQ , которая приводит A к диагональному виду. Ясно, что элементами D_i являются α_i в некотором порядке. Вспоминая, что δ_i , взятые с соответствующими кратностями, суть γ_i , имеем

$$\min_R f(R) = \sum (\gamma_i - \alpha_{p_i})^2, \quad (51.5)$$

где $p_1, p_2, \dots, p_n$ — перестановка $1, 2, \dots, n$.

Покажем теперь, что минимум достигается при $p_i = i$. Пусть

$$x = \sum (\gamma_i - \alpha_{p_i})^2 \quad (51.6)$$

при некоторой частной перестановке. Если $p_1 \neq 1$, предположим, что $p_s = 1$, так что γ_s в паре с α_1 . Переставим p_1 и p_s в перестановке. Тогда изменение в x будет

$$(\gamma_1 - \alpha_1)^2 + (\gamma_s - \alpha_{p_1})^2 - (\gamma_1 - \alpha_{p_1})^2 - (\gamma_s - \alpha_1)^2 = \\ = -2(\gamma_s - \gamma_1)(\alpha_{p_1} - \alpha_1) \leq 0, \quad (51.7)$$

и, следовательно, сумма уменьшится. Аналогично, сохраняя α_1 на первом месте и связывая вместе γ_2 и α_2 , мы видим снова, что сумма не возрастает. Следовательно, если каждое α_i вычитается из соответствующего γ_i , то сумма не больше, чем ее исходное значение.

Наш последний результат состоит в том, что минимальное значение $f(R)$ равно $\sum (\gamma_i - \alpha_i)^2$. Однако в (49.3) мы показали, что $\sum \beta_i^2$ — одно из допустимых значений для $f(R)$, и, следовательно,

$$\|B\|_E^2 = \sum_{i=1}^n \beta_i^2 \geq \sum_{i=1}^n (\gamma_i - \alpha_i)^2. \quad (51.8)$$

Это завершает доказательство.

52. Наше доказательство тривиально распространяется на эрмитовы матрицы, но результат справедлив также для любой нормальной матрицы. В общем случае могут быть использованы сходные методы доказательства, хотя теперь элементы в (p, q) - и (q, p) -позициях в (50.4) не будут комплексно сопряженными. Используя матрицы плоских вращений (43.4) главы 1, можем показать, что минимум достигается при диагональной $R^H A R$. Детали оставляются в качестве упражнения.

Дополнительные замечания

Хотя материал §§ 5—12 — в основном классическая теория возмущений, его простое, но строгое изложение трудно найти в литературе.

Теорема Гершгорина широко рекомендовалась для локализации собственных значений; особенно пылким ее сторонником была О. Таусски. Ею даны интересные обсуждения этой теоремы и ее обобщения (1949). Использование диагональных преобразований подобия для уточнения локализации собственных значений обсуждалось в оригинальной работе Гершгорина (1931), но расширение этого метода для получения результатов классической теории возмущений и строгой оценки ошибок ново.

Всеобъемлющее обсуждение включающих и исключающих областей для собственных значений, которые можно вывести применением норм, дано в главе 3 книги Хаусхолдера «The theory of matrices in numerical analysis» (1964). Эта глава содержит также обсуждение результатов, полученных из расширения принципа минимакса.

Теорема Виландта — Гофмана, кажется, не привлекает столько внимания, как теоремы, полученные прямым применением норм. По моему опыту, это наиболее полезный результат при анализе ошибок методов, основанных на ортогональных преобразованиях в арифметике с плавающей запятой.

АНАЛИЗ ОШИБОК

Введение

1. Так как главным предметом этой книги должна стать сравнительная оценка различных способов решения алгебраической проблемы собственных значений, мы включили анализ ошибок большого числа наиболее важных методов. Такой анализ, естественно, требует понимания основных арифметических операций, выполняемых на цифровой вычислительной машине. Уилкинсон (1963 г., гл. 1) дал довольно подробный подсчет ошибок округления в этих операциях и рассмотрел основные изменения, которые должны быть сделаны в существующих вычислительных машинах.

Я буду предполагать, что читатель хорошо знаком с этой книгой, и дам лишь краткую сводку результатов, которые нам здесь потребуются. Будет более просто, если мы ограничимся конкретным множеством процессов округления, так как оценки, которые мы получаем, не очень чувствительны к существующим различиям. Коснемся вычислений как с плавающей, так и с фиксированной запятой и на всем протяжении будем предполагать, что используется двоичная арифметика. Тем не менее время от времени полезно показывать простые примеры; почти во всех случаях они были выполнены на настоящих вычислительных машинах в десятичной системе счисления. Несколько хуже проводить такие расчеты на автоматических вычислительных машинах, так как ошибки округления могут оказаться включенными в ошибки перевода из двоичного представления в десятичное.

Операции с фиксированной запятой

2. При вычислениях с фиксированной запятой мы предположим, что в случае необходимости введен скалярный множитель, так что каждое число x находится в стандартном интервале

$$-1 \leq x \leq 1, \quad (2.1)$$

и не будем беспокоиться об исключении одной из граничных точек, так как это часто усложняет анализ ошибок в несущественной части. Предположим, что после запятой в двоичном числе имеется t разрядов, и будем говорить, что вычислительная машина имеет « t -разрядное слово», если даже для его представления может потребоваться $t + 1$ разряд.

Равенство вида

$$z = fi(x \overset{\times}{\underset{\div}{+}} y) \quad (2.2)$$

будет означать, что x , y и z — стандартные числа с фиксированной запятой и что z получено из x и y выполнением соответствующей операции с фиксированной запятой. В случае сложения, вычитания и деления предположим, что вычисленное z не лежит вне дозволенного интервала.

Ошибки округления не вызываются ни сложением, ни вычитанием, так что имеем

$$z = fi(x \pm y) \equiv x \pm y, \quad (2.3)$$

где последний член в (2.3) представляет точную сумму или разность. Мы используем символ эквивалентности, чтобы подчеркнуть, что ошибки округления были приняты в расчет. Это помогает отличить вычислительное равенство от математического, которое может быть использовано для описания алгоритма. Умножение и деление в общем случае вызывают ошибки округления. Мы предполагаем, что процесс округления таков, что

$$z = fi(x \overset{\times}{\div} y) \equiv x \overset{\times}{\div} y + \varepsilon, \quad (2.4)$$

где

$$|\varepsilon| \leq 2^{-t-1}. \quad (2.5)$$

Подчеркиваем, что в правой части (2.4) член $x \overset{\times}{\div} y$ представляет точное произведение или частное, так что символы « $\times$ » и « $\div$ » имеют свой обычный арифметический смысл.

Накопление скалярного произведения

3. На многих вычислительных машинах можно точно вычислить скалярное произведение

$$x_1 \times y_1 + x_2 \times y_2 + \dots + x_n \times y_n \quad (3.1)$$

в арифметике с фиксированной запятой без специального программирования (если только не происходит переполнение). В общем случае точное представление суммы (3.1) требует $2t$ разрядов после двоичной запятой. Если z — число, полученное точным накоплением скалярного произведения (3.1) и последующим округлением результата для того, чтобы дать стандартное число с фиксированной запятой, мы пишем

$$z = fi_2(x_1 \times y_1 + x_2 \times y_2 + \dots + x_n \times y_n), \quad (3.2)$$

где индекс «2» означает, что было использовано точное накопление. С нашим способом округления имеем

$$z = fi_2(x_1 \times y_1 + x_2 \times y_2 + \dots + x_n \times y_n) \equiv \sum_{i=1}^n x_i y_i + \varepsilon, \quad (3.3)$$

где

$$|\varepsilon| \leq 2^{-t-1}, \quad (3.4)$$

тогда как

$$fi(x_1 \times y_1 + x_2 \times y_2 + \dots + x_n \times y_n) \equiv \sum_{i=1}^n x_i y_i + \varepsilon, \quad (3.5)$$

где

$$|\varepsilon| \leq n \cdot 2^{-t-1}. \quad (3.6)$$

Вычислительные машины, имеющие устройство для накопления скалярного произведения, обычно рассчитаны на то, чтобы допускать делимое с $2t$ разрядами после двоичной запятой, делить его на стандартное число с фиксированной запятой и давать округленное частное с фиксированной

запятой. Мы будем писать

$$w = fi_2[(x_1 \times y_1 + x_2 \times y_2 + \dots + x_n \times y_n)/z] \equiv [(\sum_{i=1}^n x_i y_i)/z] + \varepsilon, \quad (3.7)$$

где

$$|\varepsilon| \leq 2^{-t-1}. \quad (3.8)$$

Отсюда

$$|wz - \sum_{i=1}^n x_i y_i| \equiv |z\varepsilon| \leq 2^{-t-1}|z|. \quad (3.9)$$

Мы будем очень широко использовать обозначения $fi()$ и $fi_2()$. Например, если мы имеем

$$C = fi_2(A \times B), \quad (3.10)$$

где A, B, C — матрицы, то подразумеваем, что все элементы A, B, C — стандартные числа с фиксированной запятой и что каждое c_{ij} определено равенством

$$c_{ij} = fi_2(a_{i1} \times b_{1j} + a_{i2} \times b_{2j} + \dots + a_{in} \times b_{nj}). \quad (3.11)$$

В том случае, когда C получена без переполнения, мы имеем

$$\left. \begin{aligned} C = fi_2(A \times B) &\equiv AB + F, \\ |f_{ij}| &\leq 2^{-t-1}, \end{aligned} \right\} \quad (3.12)$$

где AB означает точное матричное произведение.

Следует заметить, что если выражение в $fi()$ внутри скобок очень сложное, то результат может зависеть от порядка, в котором выполняются операции, и тогда он должен быть точно определен.

Операции с плавающей запятой

4. При вычислениях с плавающей запятой каждое стандартное число x представлено упорядоченной парой чисел a и b , так что $x = 2^b a$. Здесь b — целое, положительное или отрицательное, и a — удовлетворяющее неравенствам

$$-\frac{1}{2} \geq a \geq -1 \quad \text{или} \quad \frac{1}{2} \leq a \leq 1. \quad (4.1)$$

Мы называем b *порядком* и a *мантиссой*. Будем предполагать, что a имеет t цифр после двоичной запятой, и скажем, что вычислительная машина имеет « t -разрядную мантиссу». Будем употреблять тот же символ « t », что и в вычислении с фиксированной запятой, хотя на практике в некоторых машинах, которые могут работать в обоих режимах, число разрядов у a и b вместе такое же, как у числа с фиксированной запятой. Это важно иметь в виду, если сравниваются границы ошибок, полученные при вычислениях с фиксированной запятой и с плавающей запятой. Допустим, что число нуль имеет нестандартное представление, в котором $a = b = 0$.

Равенство вида

$$z = fl\left(x \overset{+}{\underset{\div}{\times}} y\right) \quad (4.2)$$

будет означать, что x , y и z — стандартные числа с плавающей запятой и что z получено из x и y выполнением соответствующей операции с плавающей запятой. Предполагаем, что ошибки округления в этих операциях таковы, что

$$z = fl\left(x \overset{+}{\underset{\div}{\times}} y\right) \equiv \left(x \overset{+}{\underset{\div}{\times}} y\right) (1 + \varepsilon), \quad (4.3)$$

где

$$|\varepsilon| \leq 2^{-t}, \quad (4.4)$$

так что каждая из конкретных арифметических операций дает результат, имеющий незначительную ошибку округления. Как показал Уилкинсон (1963b, стр. 11), некоторые вычислительные машины обладают такими способами округления, что сложение и вычитание не обязательно дают результаты, имеющие малые относительные ошибки, однако это дает очень небольшое расхождение в оценках, полученных для обшпрных вычислений.

Упрощенные выражения для границ ошибок

5. На всем протяжении этой книги нам неоднократно будет требоваться большое число результатов и для удобства мы собираем их здесь. Они являются прямым следствием соотношений (4.3) и (4.4). Однако применение этих соотношений прежде всего приводит к оценкам вида

$$(1 - 2^{-t})^r \leq 1 + \varepsilon \leq (1 + 2^{-t})^r, \quad (5.1)$$

а они не очень удобны. Для того чтобы упростить эти оценки, сделаем предположение, что в практических приложениях, которые для нас интересны, r будет ограничено условием

$$r2^{-t} < 0,1. \quad (5.2)$$

Для любого приемлемого значения t оправданно вводить такое ограничение на r , которое вызывается у нас соображениями о величине памяти. С ограничением (5.2) имеем

$$\left. \begin{aligned} (1 + 2^{-t})^r &< 1 + (1,06) r2^{-t}, \\ (1 - 2^{-t})^r &> 1 - (1,06) r2^{-t}. \end{aligned} \right\} \quad (5.3)$$

Для того чтобы еще более упростить эти выражения, введем t_1 , определенное равенством

$$2^{-t_1} = (1,06) 2^{-t}, \quad (5.4)$$

откуда

$$t_1 = t - \log_2(1,06) = t - 0,08406, \quad (5.5)$$

так что t_1 незначительно отличается от t . Во всех дальнейших оценках будем заменять соотношение (5.1) соотношением

$$|\varepsilon| < r2^{-t_1}, \quad (5.6)$$

всякий раз, когда это выгодно, молчаливо подразумевая ограничение (5.2).

Оценки ошибок для некоторых основных вычислений с плавающей запятой

6. Сформулируем сейчас без доказательства следующие результаты:

$$(i) \quad fl(x_1 \times x_2 \times \dots \times x_n) \equiv (1 + E) \prod_{i=1}^n x_i, \quad (6.1)$$

где

$$|E| < (n-1) 2^{-t_1}. \quad (6.2)$$

$$(ii) \quad fl(x_1 + x_2 + \dots + x_n) \equiv x_1(1 + \varepsilon_1) + x_2(1 + \varepsilon_2) + \dots + x_n(1 + \varepsilon_n), \quad (6.3)$$

где

$$|\varepsilon_1| < (n-1) 2^{-t_1}, \quad |\varepsilon_r| < (n+1-r) 2^{-t_1} \quad (r=2, \dots, n). \quad (6.4)$$

Здесь предположено, что операции выполняются в таком порядке:

$$s_2 = fl(x_1 + x_2), \quad s_r = fl(s_{r-1} + x_r) \quad (r=3, \dots, n). \\ (iii) \quad fl(x^T y) = fl(x_1 y_1 + x_2 y_2 + \dots + x_n y_n) \equiv \\ \equiv x_1 y_1 (1 + \varepsilon_1) + x_2 y_2 (1 + \varepsilon_2) + \dots + x_n y_n (1 + \varepsilon_n), \quad (6.5)$$

где

$$|\varepsilon_1| < n 2^{-t_1}, \quad |\varepsilon_r| < (n+2-r) 2^{-t_1} \quad (r=2, \dots, n). \quad (6.6)$$

Мы можем написать

$$fl(x^T y) - x^T y = x^T z, \quad (6.7)$$

где

$$z = \text{diag}(\varepsilon_i) y. \quad (6.8)$$

Полезным следствием этого основного результата является

$$|fl(x^T y) - x^T y| \leq \sum |x_i| |y_i| |\varepsilon_i| \leq \quad (6.9)$$

$$\leq n 2^{-t_1} |x|^T |y| \leq \quad (6.10)$$

$$\leq n 2^{-t_1} \|x\|_2 \|y\|_2. \quad (6.11)$$

$$(iv) \quad C = fl(A + B) \equiv A + B + F, \quad (6.12)$$

где

$$c_{ij} = (a_{ij} + b_{ij})(1 + \varepsilon_{ij}), \quad |\varepsilon_{ij}| \leq 2^{-t}. \quad (6.13)$$

Следовательно,

$$|f_{ij}| = |\varepsilon_{ij}| |a_{ij} + b_{ij}|. \quad (6.14)$$

Поэтому мы имеем

$$|fl(A + B) - (A + B)| \leq 2^{-t} |A + B|. \quad (6.15)$$

Этот результат верен только для того специального процесса округления, который мы выбрали, и далек от верного для других процессов. Он нам будет не очень полезен. Однако (6.15) несомненно означает, что

$$|fl(A + B) - (A + B)| \leq 2^{-t} (|A| + |B|), \quad (6.16)$$

а это непосредственно распространяется на все процессы округления, которые я рассмотрел, если только в эту оценку включить еще и дополнительный множитель, например, 4

$$(v) \quad fl(AB) \equiv AB + F, \quad (6.17)$$

где

$$f_{ij} = a_{i1} b_{1j} \varepsilon_1^{(ij)} + a_{i2} b_{2j} \varepsilon_2^{(ij)} + \dots + a_{in} b_{nj} \varepsilon_n^{(ij)} \quad (6.18)$$

и

$$|\varepsilon_1^{(ij)}| < n 2^{-t_1}, \quad |\varepsilon_r^{(ij)}| < (n+2-r) 2^{-t_1} \quad (r=2, \dots, n). \quad (6.19)$$

Отсюда

$$|F| < |A| |D| |B| < n2^{-t_1} |A| |B|, \quad (6.20)$$

где D есть диагональная матрица, имеющая элементы

$$[n2^{-t_1}, n2^{-t_1}, (n-1)2^{-t_1}, \dots, 3 \cdot 2^{-t_1}, 2 \cdot 2^{-t_1}]. \quad (6.21)$$

$$(vi) \quad fl(\alpha A) \equiv \alpha A + F, \quad (6.22)$$

где

$$f_{ij} = \alpha a_{ij} \varepsilon_{ij} \quad \text{и} \quad |\varepsilon_{ij}| \leq 2^{-t}. \quad (6.23)$$

Следовательно,

$$|fl(\alpha A) - \alpha A| \leq 2^{-t} |\alpha| |A|. \quad (6.24)$$

$$(vii) \quad fl(xy^T) \equiv xy^T + F, \quad (6.25)$$

где

$$f_{ij} = x_i y_j \varepsilon_{ij} \quad \text{и} \quad |\varepsilon_{ij}| \leq 2^{-t}. \quad (6.26)$$

Отсюда

$$|fl(xy^T) - xy^T| \leq 2^{-t} |x| |y|^T. \quad (6.27)$$

Оценки норм матриц ошибок

7. В приложениях результатов, которые мы только что получили, нам часто будут нужны оценки для какой-нибудь нормы матрицы ошибок. Если, например, мы возьмем оценку (6.20) ошибки в матрице произведения, то

$$\|F\| \leq \|F\| \leq n2^{-t_1} \|A\| \|B\|, \quad (7.1)$$

и, следовательно, для евклидовой, 1-нормы и ∞ -нормы мы имеем

$$\|F\| \leq n2^{-t_1} \|A\| \|B\|, \quad (7.2)$$

но для 2-нормы только

$$\|F\|_2 \leq n^2 2^{-t_1} \|A\|_2 \|B\|_2. \quad (7.3)$$

В действительности оценка в (7.3) слаба в сравнении с

$$\|F\|_E \leq n2^{-t_1} \|A\|_E \|B\|_E, \quad (7.4)$$

так как

$$\|F\|_2 \leq \|F\|_E \quad \text{и} \quad \|A\|_E \|B\|_E \leq n \|A\|_2 \|B\|_2. \quad (7.5)$$

Мы часто будем убеждаться, что оценки, полученные с использованием 2-нормы, слабее, чем оценки, полученные с использованием евклидовой нормы. Заметим, однако, что если A и B имеют неотрицательные элементы, то

$$\|F\|_2 \leq n2^{-t_1} \|A\|_2 \|B\|_2, \quad (7.6)$$

и этот результат сильнее, чем результат из (7.4).

Накопление скалярных произведений в арифметике с плавающей запятой

8. На некоторых вычислительных машинах, которые сейчас строятся, предусмотрены специальные устройства для сложения и вычитания чисел, имеющих мантиссу с $2t$ двоичными разрядами. На таких машинах сумма большого числа слагаемых $fl(x_1 + x_2 + \dots + x_n)$ может быть накоплена

с $2t$ -разрядной мантиссой, так же как и скалярное произведение $fl(x_1y_1 + x_2y_2 + \dots + x_ny_n)$. В последнем предполагается, что x_i и y_i — стандартные числа с t -разрядной мантиссой. Аналогично делимое может допускаться с $2t$ -разрядной мантиссой и оно может быть разделено на стандартное число с t -разрядной мантиссой. На АСЕ, одной из вычислительных машин в Национальной физической лаборатории, существуют подпрограммы, которые выполняют эти операции почти за то же время, что и соответствующие стандартные операции с плавающей запятой. Мы используем обозначение $fl_2(\quad)$, чтобы указать на аналогию этих операций с использованными нами $fi_2(\quad)$.

Заметим, однако, что $fl_2(x_1 + x_2 + \dots + x_n)$, вообще говоря, не дает точной суммы.

Ошибки округления в основных операциях $fl_2(\quad)$ были исследованы Уилкинсоном (1963b, стр. 23—25), и мы не будем здесь повторять анализ, но приведем наиболее полезные результаты. Прямое вычисление оценок ошибок для основных арифметических операций приводит к границам вида

$$\left(1 - \frac{3}{2} 2^{-2t}\right)^r \leq 1 + \varepsilon \leq \left(1 + \frac{3}{2} 2^{-2t}\right)^r, \quad (8.1)$$

а они несколько неудобны. Мы будем всегда предполагать, что

$$\frac{3}{2} r 2^{-2t} < 0,1, \quad (8.2)$$

и с этим ограничением (8.1) означает, что

$$|\varepsilon| < \frac{3}{2} (1,06) r 2^{-2t}. \quad (8.3)$$

Соответственно определению t_1 в § 5 мы определим t_2 так, что

$$2^{-2t_2} = (1,06) 2^{-2t}, \quad 2t_2 = 2t - 0,08406. \quad (8.4)$$

Полученное t_2 лишь немного отличается от t , и (8.1) теперь означает, что

$$|\varepsilon| < \frac{3}{2} r 2^{-2t_2}. \quad (8.5)$$

Если в одном и том же вычислении используется как $fl(\quad)$, так и $fl_2(\quad)$, предположение (5.2) делает (8.2) излишним.

Оценки ошибок для некоторых основных вычислений $fl_2(\quad)$ с двойным числом разрядов

9. Мы сформулируем сейчас без доказательства следующие основные результаты:

$$(i) \quad fl_2(x_1 + x_2 + \dots + x_n) \equiv [x_1(1 + \varepsilon_1) + x_2(1 + \varepsilon_2) + \dots + x_n(1 + \varepsilon_n)](1 + \varepsilon), \quad (9.1)$$

$$\left. \begin{aligned} |\varepsilon| &\leq 2^{-t}, \quad |\varepsilon_1| < \frac{3}{2} (n-1) 2^{-2t_2}, \\ |\varepsilon_r| &< \frac{3}{2} (n+1-r) 2^{-2t_2} \quad (r = 2, \dots, n). \end{aligned} \right\} \quad (9.2)$$

В (9.1) сумма накоплена с использованием $2t$ -разрядной мантиссы и после этого округлена. Далее мы имеем

$$fl_2(x_1 + x_2 + \dots + x_n) - (x_1 + x_2 + \dots + x_n)(1 + \varepsilon) \equiv x_1\varepsilon_1 + x_2\varepsilon_2 + \dots + x_n\varepsilon_n, \quad (9.3)$$

где границы (9.2) по-прежнему имеют место.

$$(ii) \quad fl_2(x_1y_1 + x_2y_2 + \dots + x_ny_n) \equiv [x_1y_1(1 + \varepsilon_1) + x_2y_2(1 + \varepsilon_2) + \dots + x_ny_n(1 + \varepsilon_n)](1 + \varepsilon), \quad (9.4)$$

$$\left. \begin{aligned} |\varepsilon| &\leq 2^{-t}, \quad |\varepsilon_1| < \frac{3}{2} n 2^{-2t_2}, \\ |\varepsilon_r| &< \frac{3}{2} (n + 2 - r) 2^{-2t_2} \quad (r = 2, \dots, n). \end{aligned} \right\} \quad (9.5)$$

Окончательно

$$fl_2(x^Ty) - x^Ty(1 + \varepsilon) \equiv x_1y_1\varepsilon_1 + x_2y_2\varepsilon_2 + \dots + x_ny_n\varepsilon_n, \quad (9.6)$$

где все еще имеют место оценки (9.5). Важно понять все значение (9.6). Оно не дает нам возможности утверждать, что $fl_2(x^Ty)$ имеет незначительную *относительную ошибку*, так как среди членов x^Ty может иметь место точное сокращение. Однако (9.6) означает, что

$$\begin{aligned} |fl_2(x^Ty) - x^Ty| &\leq |\varepsilon| |x^Ty| + \sum |x_i| |y_i| |\varepsilon_i| \leq \\ &\leq 2^{-t} |x^Ty| + \frac{3}{2} n 2^{-2t_2} |x|^T |y| \leq \end{aligned} \quad (9.7)$$

$$\leq 2^{-t} |x^Ty| + \frac{3}{2} n 2^{-2t_2} \|x\|_2 \|y\|_2. \quad (9.8)$$

Если среди членов x^Ty точное сокращение не имеет места, то второй член в правой части (9.7) пренебрежимо мал по сравнению с первым.

(iii) Полагая в результате (ii) y равным x , имеем

$$x^Tx = |x|^T |x|, \quad (9.9)$$

и, следовательно,

$$|fl_2(x^Tx) - x^Tx| \leq 2^{-t} (x^Tx) + \frac{3}{2} n 2^{-2t_2} (x^Tx) = \left(2^{-t} + \frac{3}{2} n 2^{-2t_2}\right) (x^Tx). \quad (9.10)$$

Если $\frac{3}{2} n 2^{-t} < 0,1$, то

$$2^{-t} + \frac{3}{2} n 2^{-2t_2} < 2^{-t} [1 + (0,1) 2^{2t-2t_2}] < 2^{-t} (1,11), \quad (9.11)$$

так что $fl_2(x^Tx)$ всегда имеет незначительную относительную ошибку.

(iv) Если x и y векторы, а z — число, то

$$fl_2\left(\frac{x^Ty}{z}\right) \equiv \frac{x_1y_1(1 + \varepsilon_1) + \dots + x_ny_n(1 + \varepsilon_n)}{z/(1 + \varepsilon)}, \quad (9.12)$$

где ε и ε_i удовлетворяют соотношениям (9.5). Мы можем это выразить в виде: вычисленное значение x^Ty/z равно точному значению $\overline{x^Ty/z}$, где элементы $\overline{x}$ и $\overline{y}$ отличаются от элементов x и y относительными ошибками

порядка 2^{-2t_2} , тогда как $\bar{z}$ отличается от z относительной ошибкой порядка 2^{-t} .

$$(v) \quad fl_2(AB) \equiv AB + F, \quad (9.13)$$

где

$$|F| \leq 2^{-t} |AB| + \frac{3}{2} n 2^{-2t_2} |A| |B|. \quad (9.14)$$

Вычисление квадратных корней

10. Квадратные корни обычно неизбежны на том или ином этапе в методах, включающих использование элементарных унитарных преобразований. Оценка ошибки, сделанной при извлечении квадратного корня, естественно, зависит до некоторой степени от того алгоритма, который используется. Мы не хотим вступать в детальное обсуждение таких алгоритмов. Будем предполагать, что

$$fi(x^{1/2}) \equiv x^{1/2} + \varepsilon, \quad \text{где} \quad |\varepsilon| < (1,00004) 2^{-t-1}, \quad (10.1)$$

$$fl(x^{1/2}) \equiv x^{1/2}(1 + \varepsilon), \quad \text{где} \quad |\varepsilon| < (1,00004) 2^{-t}, \quad (10.2)$$

так как это верно на АСЕ. В большинстве матричных алгоритмов число извлечений квадратных корней мало по сравнению с числом других операций, и даже существенно больше ошибки, чем те, что в (10.1) и (10.2), не смогут внести сколько-нибудь значительной разницы в полные оценки ошибок.

Блочно-плавающие векторы и матрицы

11. В арифметике с фиксированной запятой общим принципом является использование такого метода представления векторов и матриц, который до некоторой степени сочетает достоинства фиксированной и плавающей запятой.

В этой схеме всем компонентам вектора или матрицы ставится в соответствие единственный числовой множитель в виде степени 2. Этот множитель выбирается так, чтобы модули самых больших компонент лежали между $1/2$ и 1 . Такой вектор называется *нормализованным блочно-плавающим вектором*. Так, например, в (11.1). a есть нормализованный блочно-плавающий вектор, а A — *нормализованная блочно-плавающая матрица* (мы, естественно, использовали для иллюстрации десятичную систему счисления):

$$a = 10^{-4} \begin{bmatrix} 0,0013 \\ -0,2763 \\ 0,0002 \\ 0,0013 \end{bmatrix}, \quad A = 10^5 \begin{bmatrix} 0,0067 & 0,2175 & 0,4132 \\ -0,1352 & 0,3145 & -0,5173 \\ -0,0167 & -0,0004 & 0,5432 \end{bmatrix}. \quad (11.1)$$

Иногда столбцы матрицы настолько различаются по величине, что каждый столбец представляется как блочно-плавающий вектор. Аналогично и каждую строку мы можем представить как блочно-плавающий вектор. Наконец, мы могли бы получить скалярный множитель, единый для всех строк и столбцов.

Преимущество этой схемы состоит в том, что только одно слово представляет собой порядок, а каждый отдельный элемент запоминается пол-

ным словом без затраты разрядов на порядок, как это было бы в вычислении с плавающей запятой.

Часто на промежуточной стадии вычислений вычисляется такой блочно-плавающий вектор, у которого нет компонент, по модулю больших $1/2$. Такой вектор называется *ненормализованным* блочно-плавающим вектором и иллюстрируется вектором a в (11.2)

$$a = 10^3 \begin{bmatrix} -0,0023 \\ 0,0123 \\ 0,0003 \\ -0,0021 \end{bmatrix}. \quad (11.2)$$

В некоторых случаях числовой множитель, соответствующий вектору, не имеет никакого значения; это верно, например, для собственного вектора. Мы определяем *нормированный блочно-плавающий вектор* как вектор, который должен быть нормализованным блочно-плавающим вектором, имеющим числовой множитель 2^0 . В таком векторе числовой множитель обычно опускается.

Основные ограничения t -разрядных вычислений

12. В главе 2 мы уже отмечали, что если матрица A требует для своего представления больше чем t разрядов, то мы не можем ее представить в вычислительной машине, имеющей t -разрядное слово. Мы должны с самого начала ограничиться « t -разрядной аппроксимацией матрицы».

Допустим, что мы согласились на момент считать исходную матрицу точно записанной в вычислительной машине; тогда мы могли бы спросить себя, какова оптимальная точность, которую можем разумно ожидать от любого метода, если используем, скажем, плавающую арифметику с t -разрядной мантиссой. (В дальнейшем будем говорить об этом как о *t -разрядной арифметике с плавающей запятой*). Мы можем пролить некоторый свет на этот вопрос, если рассмотрим очень простую операцию над A — вычисление преобразования $D^{-1}AD$, где $D = \text{diag}(d_i)$. Если напомним

$$B = fl(D^{-1}AD), \quad (12.1)$$

то увидим, что

$$b_{ij} = d_i^{-1} a_{ij} d_j (1 + \varepsilon_{ij}), \quad (12.2)$$

где

$$|\varepsilon_{ij}| < 2 \cdot 2^{-t_i}. \quad (12.3)$$

Соотношения (12.2) показывают, что B есть точное подобное преобразование матрицы с элементами $a_{ij}(1 + \varepsilon_{ij})$, и, следовательно, собственные значения у B те же, что у матрицы $(A + F)$, где

$$|F| < 2 \cdot 2^{-t_i} |A|. \quad (12.4)$$

Нетрудно построить примеры, для которых эти границы почти достижимы.

Поэтому собственные значения λ'_i матрицы B будут до некоторой степени отличаться от собственных значений λ_i матрицы A . Для изолированных собственных значений мы знаем из гл. 2, § 9, что при $t \rightarrow \infty$

$$\lambda'_i - \lambda_i \approx y_i^T F x_i / s_i, \quad (12.5)$$

где y_i и x_i — нормированные левые и правые собственные векторы A . Из (12.4) имеем

$$|y_i^T F x_i / s_i| < 2 \cdot 2^{-t_1} \|A\|_E / |s_i| \leq 2 \cdot 2^{-t_1} n^{1/2} \|A\|_2 / |s_i|. \quad (12.6)$$

Отсюда выводим, что даже простейшие подобные преобразования приводят к ошибке в собственном значении λ_i , которая может считаться пропорциональной 2^{-t} и $1/|s_i|$. Если $1/|s_i|$ велико, то весьма правдоподобно, что собственное значение λ_i будет изменяться в соответствующей степени в зависимости от ошибок округления, сделанных при преобразовании. Заметим, кроме этого, что если исходная матрица не была задана точно, то ошибки такого порядка уже были введены.

Большое число методов определения собственных значений включает в себя вычисление значительного числа подобных преобразований, каждое из которых существенно сложнее, чем то, которое мы только что рассмотрели. Предположим, что есть k таких преобразований. Кажется совсем невероятным, что мы сможем получить априорную оценку для ошибки, которая была бы меньше k -кратной величины только что рассмотренной. Если преобразование было в некотором смысле неустойчивым, то мы могли бы ожидать, что будут иметь место и много большие ошибки.

Если метод включает k подобных преобразований и мы можем доказать, что конечная матрица точно подобна $(A + F)$ и можем получить оценку вида

$$|F| < 2k \cdot 2^{-t_1} |A|, \quad (12.7)$$

которая справедлива для всех A , то будем говорить, что такой численный метод должен рассматриваться как необычайно устойчивый. Необходимо подчеркнуть, что даже такой метод, как этот, может дать неточные результаты для тех собственных значений, для которых соответствующие значения $1/|s_i|$ велики. Мы рассматриваем любую ошибку, которая возникает из-за этого фактора, как неизбежную и никоим образом не должны приписывать ее недостаткам рассматриваемого метода.

Читатель может почувствовать, что на этом этапе у нас мало оправданий по замечаниям последнего параграфа. Мы не будем пытаться защищать их здесь, но надеемся, что анализ, данный в последующих главах, убедит читателя в их справедливости. Однако, чтобы избежать совсем неправильного понимания нашей точки зрения, мы добавим несколько замечаний.

13. (i). Мы не предполагаем, что если некоторая матрица имеет большое значение какого-нибудь из чисел $1/s_i$, то соответствующее собственное значение должно *обязательно* иметь потерю точности порядка величины, даваемой оценками в (12.5) и (12.6), при любом используемом методе. Рассмотрим, например, плохо обусловленную матрицу, определенную в § 33 гл. 2; она принадлежит классу матриц, которые называются *трехдиагональными*. Матрица A является трехдиагональной, если

$$a_{ij} = 0 \quad \text{при} \quad |j - i| > 1 \quad (13.1)$$

Теперь будем развивать методы, предназначенные специально для трехдиагональных матриц, и покажем, что ошибки округления, сделанные в этих методах, эквивалентны возмущениям в *ненулевых элементах* A . Хотя матрица в § 33 имеет плохо обусловленные собственные значения, они не очень чувствительны к возмущениям элементов в трех центральных линиях по диагонали, и, следовательно, используя эти методы, мы

вполне можем получить для данной матрицы очень точные результаты, несмотря на большие значения $1/|s_i|$.

(ii) Если мы накапливаем скалярное произведение или используем какие-либо операции типа fi_2 () или fl_2 (), то хотя можно представить это себе как t -разрядную арифметику, мы эффективно получаем некоторые результаты, которые имеют $2t$ верных цифр. Если такие операции используются широко, мы можем получить заметно лучшие результаты.

(iii) В обсуждении предыдущего раздела мы ссылались на возможность того, что некоторые преобразования могли быть «неустойчивыми», и подразумевали, что тогда ошибки могли бы быть существенно больше. Мы можем проиллюстрировать эту идею неустойчивости очень простым способом. Пусть матрица A_0 определена так:

$$A_0 = \begin{bmatrix} 10^{-3}(0,3000) & 0,4537 \\ 0,3563 & 0,5499 \end{bmatrix}. \quad (13.2)$$

Используя 4-разрядную плавающую арифметику, умножим A_0 справа на матрицу типа M_1 (гл. 1, § 40), выбирая ее так, чтобы сделать элемент (1, 2) нулем. Мы имеем

$$M_1 = \begin{bmatrix} 1 & 10^4(-0,1512) \\ 0 & 1 \end{bmatrix},$$

$$A_0 M_1 = \begin{bmatrix} 10^{-3}(0,3000) & 0 \\ 0,3563 & 10^3(-0,5382) \end{bmatrix} \quad (13.3)$$

(где в качестве элемента (1, 2) матрицы $A_0 M_1$ написан 0). Далее

$$M_1^{-1} = \begin{bmatrix} 1 & 10^4(0,1512) \\ 0 & 1 \end{bmatrix},$$

$$M_1^{-1} A_0 M_1 = \begin{bmatrix} 10^3(0,5387) & 10^6(-0,8138) \\ 0,3563 & 10^3(-0,5382) \end{bmatrix} = A_1. \quad (13.4)$$

Так как след вычисленной матрицы равен 0,5000, а исходной матрицы равен 0,5502, то A_1 не может быть точно подобной любой $(A_0 + F)$, для которой

$$|F| < 2 \cdot 10^{-4} |A_0|. \quad (13.5)$$

Очень большой элемент в M_1 привел к A_1 , имеющей много большие элементы, чем A_0 , и ошибки округления не эквивалентны «малым» возмущениям в элементах A_0 . В действительности пагубный эффект преобразования более сильный, чем может быть замечен просто из рассмотрения ошибки в следе. Собственные значения A_0 с четырьмя десятичными знаками суть 0,7621 и $-0,2119$, в то время как у A_1 они равны $0,2500 \pm \pm i(5,342)$. В общем случае мы убедимся в том, что при использовании матриц преобразования с большими элементами ошибки округления эквивалентны большим возмущениям в исходной матрице.

Методы определения собственных значений, основанные на подобных преобразованиях

14. Учитывая соображения предыдущего раздела, даем общий анализ методов определения собственных значений, основанных на подобных преобразованиях. Пусть рассматривается алгоритм, который по матрице A_0 определяет последовательность матриц $A_1, A_2, \dots, A_s$, где

$$A_p = H_p^{-1} A_{p-1} H_p \quad (p = 1, \dots, s). \quad (14.1)$$

Обычно, но не всегда, каждая A_p более проста, чем ее предшественница, в том смысле, что H_p определяется таким способом, при котором A_p сохраняет все нулевые элементы, полученные на прежних шагах, и дополнительно имеет несколько нулевых элементов. Существует несколько методов итерационного характера, для которых, однако, последовательность существенно бесконечна, и нулевые элементы получаются только в предельной матрице последовательности. Методы этого типа должны обладать на практике достаточно быстрой сходимостью для элементов, чтобы сделать «нуль с рабочей точностью» за умеренное число шагов. Анализ, который дается ниже, применяется в любом случае; если метод итерационного типа, то множество нулевых элементов, на которое мы время от времени ссылаемся, может быть пустым.

Последовательность матриц A_p , определенных (14.1), соответствует, конечно, точному вычислению. В действительности делаются ошибки округления и в результате получается последовательность матриц, которые обозначим как $A_0, \bar{A}_1, \bar{A}_2, \dots, \bar{A}_p$. Матрица $\bar{A}_p$ имеет одно особое свойство, которое заслуживает замечания. Точный алгоритм конструируется так, чтобы получить некоторое количество нулевых элементов в каждой A_p ; в действительном процессе матрицам $\bar{A}_p$ также приписывается то же самое множество нулевых элементов. Другими словами, не нужно пытаться вычислять эти элементы; они автоматически устанавливаются равными нулю на каждой стадии процесса. Рассмотрим теперь типичную стадию в практической процедуре, когда мы уже получили матрицу $\bar{A}_{p-1}$. Существует матрица H'_p , отличная от матрицы H_p , которая получена применением точного алгоритма к $\bar{A}_{p-1}$. Кроме того, в действительности получаем матрицу $\bar{H}_p$, которая, вообще говоря, не будет той же, что H'_p , потому что в применении алгоритма к $\bar{A}_{p-1}$ мы будем делать ошибки округления. Таким образом, имеется три последовательности матриц:

- (i) матрицы H_p и A_p , удовлетворяющие (14.1) и соответствующие точному выполнению алгоритма;
- (ii) матрицы H'_p и A'_p , удовлетворяющие

$$A'_p = (H'_p)^{-1} \bar{A}_{p-1} H'_p, \quad (14.2)$$

где H'_p есть матрица, соответствующая точному применению p -го шага алгоритма к $\bar{A}_{p-1}$;

(iii) матрицы $\bar{H}_p$ и $\bar{A}_p$, которые получаются на практике.

Мы найдем, что первая последовательность матриц не играет никакой роли в любом анализе ошибок, который будет дан. *Можно было бы ожидать, что алгоритмы, которые численно устойчивы, приведут к матрицам $\bar{A}_p$, близким к A_p , и что будет плодотворной попыткой нахождение оценки для $\bar{A}_p - A_p$, но это неверно.* На первый взгляд это может показаться едва ли заслуживающим доверия, так как кажется, что если $\bar{A}_p$ не близка к A_p , то не должно быть причин предполагать, что собственные значения $\bar{A}_p$ близки к собственным значениям A_p . Однако мы должны помнить, что в действительности нам неинтересно, близка ли $\bar{A}_p$ к A_p , если только она точно подобна некоторой матрице, которая близка к A_p . Если это верно, то ошибка в каждом собственном значении будет зависеть главным образом от их чувствительности к возмущениям в исходной матрице, и мы уже отмечали, что не можем ожидать устранения отрицательного эффекта ее плохой обусловленности.

Анализ ошибок методов, основанных на элементарных неунитарных преобразованиях

15. Соображения, рассматриваемые в общем анализе, различны в зависимости от того, являются ли матрицы H_p унитарными или неунитарными элементарными матрицами. Мы рассмотрим сначала неунитарный случай.

На p -м шаге находим численно матрицу $\bar{H}_p$; по ней и $\bar{A}_{p-1}$ определяем матрицу $\bar{A}_p$. Все элементарные неунитарные матрицы такой природы, что, имея $\bar{H}_p$, обратная матрица $\bar{H}_p^{-1}$ находится немедленно без ошибок округления, как было показано в гл. 1 § 40. В общем случае элементы $\bar{A}_p$ распадаются на три основных класса:

(i) Элементы, которые были уже исключены на предыдущих шагах и остаются нулевыми. Даже если p -й шаг, начиная с $\bar{A}_{p-1}$, осуществлялся неточно, эти элементы будут оставаться нулевыми, так как новые элементы обычно определяются как линейные комбинации нулевых элементов $\bar{A}_{p-1}$.

(ii) Элементы, которые предназначаются по алгоритму для исключения на p -м шаге. Им автоматически приписываются нулевые значения, даже если соответствующие элементы $\bar{H}_p^{-1}\bar{A}_{p-1}\bar{H}_p$ не точно равны нулю из-за ошибок округления, сделанных при вычислении $\bar{H}_p$. Заметим, что элементы $(H'_p)^{-1}\bar{A}_{p-1}H'_p$ в этих позициях точно равны нулю, так как по определению H'_p есть матрица, соответствующая точному применению p -го шага алгоритма к $\bar{A}_{p-1}$.

(iii) Элементы, которые получаются вычислением $\bar{H}_p^{-1}\bar{A}_{p-1}\bar{H}_p$. Определим матрицу F_p соотношением

$$\bar{A}_p \equiv \bar{H}_p^{-1}\bar{A}_{p-1}\bar{H}_p + F_p \quad (p = 1, \dots, s), \quad (15.1)$$

так что она означает разность между матрицей, которую мы принимаем за $\bar{A}_p$, и точным произведением $\bar{H}_p^{-1}\bar{A}_{p-1}\bar{H}_p$.

Уравнения (15.1) могут быть скомбинированы так:

$$\bar{A}_s \equiv F_s + G_s^{-1}F_{s-1}G_s + G_{s-1}^{-1}F_{s-2}G_{s-1} + \dots + G_2^{-1}F_1G_2 + G_1^{-1}A_0G_1, \quad (15.2)$$

где

$$G_p = \bar{H}_p\bar{H}_{p+1}\dots\bar{H}_s. \quad (15.3)$$

Если мы определим F равенством

$$F = F_s + G_s^{-1}F_{s-1}G_s + G_{s-1}^{-1}F_{s-2}G_{s-1} + \dots + G_2^{-1}F_1G_2, \quad (15.4)$$

то (15.2) дает

$$\bar{A}_s \equiv F + G_1^{-1}A_0G_1, \quad \text{или} \quad \bar{A}_s - F \equiv G_1^{-1}A_0G_1. \quad (15.5)$$

Это означает, что $\bar{A}_s$ отличается от точного подобного преобразования A_0 на матрицу F . С другой стороны, если мы определим K равенством

$$K = G_1FG_1^{-1}, \quad (15.6)$$

то (15.2) дает

$$\bar{A}_s \equiv G_1^{-1}(K + A_0)G_1. \quad (15.7)$$

Это показывает, что $\bar{A}_s$ точно подобна $(A_0 + K)$ и, следовательно, мы получили выражение для возмущения в A_0 , которое эквивалентно ошибкам, сделанным при вычислении. Из (15.3), (15.4) и (15.6) имеем

$$K = L_sF_sL_s^{-1} + L_{s-1}F_{s-1}L_{s-1}^{-1} + \dots + L_1F_1L_1^{-1}, \quad (15.8)$$

где

$$L_p = \bar{H}_1 \bar{H}_2 \dots \bar{H}_p. \quad (15.9)$$

Теперь, учитывая то, что мы говорили ранее, можно считать отношение $\|K\|/\|A_0\|$ как меру эффективности рассматриваемого алгоритма. Чем меньше априорные оценки мы сможем найти для этого отношения, тем более эффективным нужно считать алгоритм.

Мы увидим, что чрезвычайно трудно найти реально пригодные априорные оценки матриц возмущения K для методов, основанных на неунитарных элементарных преобразованиях. Однако уравнение (15.8) дает строгое указание на то, что желательно использовать такие элементарные матрицы, которые не имеют большой нормы. В большинстве случаев будем использовать матрицы типа M_p или N_p из гл. 1, § 40 и в интересах численной устойчивости обычно будем строить наши алгоритмы так, чтобы элементы M_p и N_p были ограничены по модулю единицей, если это возможно.

Этот вывод усиливается, если мы заметим, что преобразование с матрицей $\bar{H}_p$, имеющей большие элементы, вероятно, должно привести к $\bar{A}_p$ с много большими элементами, чем у $\bar{A}_{p-1}$, и поэтому ошибки округления на этом этапе, вероятно, должны быть соответственно большими, приводя к F_p с большой нормой (см. пример из § 13).

Мы дали выражение для эквивалентных возмущений F и K соответственно в конечной вычисленной матрице $\bar{A}_s$ и начальной матрице A_0 . Оценки для F , вероятно, должны быть полезны *апостериорно*, когда мы вычислили собственную систему матрицы $\bar{A}_s$ (по крайней мере приближенно) и имеем оценки для обусловленности ее собственных значений. С другой стороны, *априорные* оценки для K дают возможность оценить алгоритм.

Возвращаясь теперь к отдельным уравнениям (15.1), мы видим, что если хотим проследить историю собственного значения λ_i , нам необходимо иметь оценки для $s_i^{(p)}$ на каждом этапе. Вообще говоря, $s_i^{(p)}$ будут различными для каждого значения p . Если бы мы смогли гарантировать, что обусловленность собственных значений не ухудшилась, то мы бы знали, что влияние различных F_p было не более чем аддитивное.

Анализ ошибок методов, основанных на элементарных унитарных преобразованиях

16. Вообще говоря, мы можем получить гораздо лучшие границы ошибок для методов, основанных на элементарных унитарных преобразованиях, но есть одна трудность, которая не возникает в случае неунитарных преобразований.

На p -м этапе вычисляем матрицу $\bar{H}_p$, которая, если мы не делали ошибок на этом этапе, должна была бы быть точно унитарным преобразованием H'_p , определенным по алгоритму для матрицы $\bar{A}_{p-1}$. Имея вычисленную матрицу $\bar{H}_p$, мы предполагаем, что она действительно унитарная, и берем $\bar{H}_p$ как ее обратную. Было бы *желательно*, чтобы мы делали так во всех случаях, но когда исходная матрица A_0 эрмитова, это почти обязательно, так как мы, конечно, захотим сохранить эрмитову природу $\bar{A}_p$. Однако при использовании формулы $\bar{H}_p^H \bar{A}_{p-1} \bar{H}_p$ мы даже не пытаемся вычислить подобное преобразование $\bar{A}_{p-1}$.

Важно, чтобы метод вычисления $\bar{H}_p$ был таким, чтобы он давал матрицу, которая близка к точно унитарной матрице. Обычно мы будем обеспечивать близость $\bar{H}_p$ к H'_p , точной унитарной матрице, определенной p -м шагом алгоритма для $\bar{A}_{p-1}$, но это не единственный путь, как мы и покажем в § 19.

Предположим, что

$$\bar{H}_p \equiv H'_p + X_p, \quad (16.1)$$

и мы можем найти достаточно малую оценку для X_p . Тогда определим F_p , как в (15.1), соотношением

$$\bar{A}_p \equiv \bar{H}_p^H \bar{A}_{p-1} \bar{H}_p + F_p, \quad (16.2)$$

так что F_p есть опять разность между принятой $\bar{A}_p$ и точным произведением $\bar{H}_p^H \bar{A}_{p-1} \bar{H}_p$. Теперь имеем

$$\bar{A}_p \equiv (H'_p + X_p)^H \bar{A}_{p-1} (H'_p + X_p) + F_p = (H'_p)^H \bar{A}_{p-1} H'_p + Y_p, \quad (16.3)$$

где

$$Y_p = (H'_p)^H \bar{A}_{p-1} X_p + (X_p)^H \bar{A}_{p-1} H'_p + X_p^H \bar{A}_{p-1} X_p + F_p, \quad (16.4)$$

и $(H'_p)^H \bar{A}_{p-1} H'_p$ есть точное унитарное подобное преобразование $\bar{A}_{p-1}$.

Объединяя уравнения (16.3) для значений p от 1 до s , имеем

$$\bar{A}_s \equiv Y_s + G_s^H Y_{s-1} G_s + G_{s-1}^H Y_{s-2} G_{s-1} + \dots + G_2^H Y_1 G_2 + G_1^H A_0 G_1, \quad (16.5)$$

где

$$G_p = H'_p H'_{p+1} \dots H'_s. \quad (16.6)$$

Сравнивая это равенство с (15.2) и (15.3), видим, что H'_p аналогична $\bar{H}_p$, а Y_p аналогична F_p . Мы можем написать

$$\bar{A}_s \equiv Y + G_1^H A_0 G_1, \quad (16.7)$$

где

$$Y = Y_s + G_s^H Y_{s-1} G_s + G_{s-1}^H Y_{s-2} G_{s-1} + \dots + G_2^H Y_1 G_2, \quad (16.8)$$

или же

$$\bar{A}_s \equiv G_1^H (Z + A_0) G_1, \quad (16.9)$$

где

$$Z = L_s Y_s L_s^H + L_{s-1} Y_{s-1} L_{s-1}^H + \dots + L_1 Y_1 L_1^H \quad (16.10)$$

и

$$L_p = H'_1 H'_2 \dots H'_p. \quad (16.11)$$

Преимущество унитарных преобразований

17. Теперь, несмотря на большую сложность определения Y_p по сравнению с определением F_p в § 15, уравнения от (16.5) до (16.11) несут гораздо больше информации, чем соответствующие уравнения § 15. Все матрицы G_p и L_p точно унитарны, и, следовательно, мы имеем

$$\|Y\|_2 \leq \|Y_s\|_2 + \|Y_{s-1}\|_2 + \dots + \|Y_1\|_2, \quad (17.1)$$

учитывая инвариантность 2-нормы к унитарным преобразованиям. Аналогично, так как $Z = G_1 Y G_1^H$, имеем

$$\|Z\|_2 = \|Y\|_2. \quad (17.2)$$

Уравнение (16.7) показывает, что $(\bar{A}_s - Y)$ — точное унитарное подобное преобразование A_0 , тогда как (16.9) показывает, что $\bar{A}_s$ точное унитарное подобное преобразование матрицы $(A_0 + Z)$.

На практике мы будем использовать или плоские вращения, или матрицы отражения и найдем, что во всех случаях сможем получить достаточно удовлетворительные оценки для нормы матрицы X_p , определенной (16.1). Если мы сможем доказать, например, что

$$\|X_p\|_2 \leq a2^{-t}, \quad (17.3)$$

то (16.4) дает

$$\|Y_p\|_2 \leq \|\bar{A}_{p-1}\|_2(2a2^{-t} + a^22^{-2t}) + \|F_p\|_2. \quad (17.4)$$

Из (16.3)

$$\|\bar{A}_p\|_2 \leq \|\bar{A}_{p-1}\|_2 + \|Y_p\|_2 \leq (1 + a2^{-t})^2 \|\bar{A}_{p-1}\|_2 + \|F_p\|_2. \quad (17.5)$$

Здесь F_p есть матрица ошибок, полученных при попытке вычислить $\bar{H}_p^H \bar{A}_{p-1} \bar{H}_p$, и так как $\bar{H}_p$ — почти унитарная матрица, общий порядок величины элементов $\bar{A}_p$ будет почти таким же, как у элементов $\bar{A}_{p-1}$. Поэтому на практике обычно нетрудно найти оценку для $\|F_p\|_2$ вида

$$\|F_p\|_2 \leq f(p, n) 2^{-t} \|\bar{A}_{p-1}\|_2, \quad (17.6)$$

где $f(p, n)$ — некоторая простая функция от p и n . Из (17.4), (17.5), (17.6) имеем

$$\|\bar{A}_p\|_2 \leq [(1 + a2^{-t})^2 + f(p, n) 2^{-t}] \|\bar{A}_{p-1}\|_2, \quad (17.7)$$

и

$$\|Y_p\|_2 \leq [2a + a^22^{-t} + f(p, n)] 2^{-t} \|\bar{A}_{p-1}\|_2. \quad (17.8)$$

Отсюда мы можем получить априорные оценки для нормы эквивалентного возмущения Z в A_0 в виде

$$\|Z\|_2 \leq 2^{-t} \|A_0\|_2 \sum_{p=1}^s x_p, \quad (17.9)$$

где

$$x_p = [2a + a^22^{-t} + f(p, n)] y_p, \quad (17.10)$$

$$y_p = \prod_{i=1}^p (1 + a2^{-t})^2 + f(i, n) 2^{-t}. \quad (17.11)$$

Анализ любого алгоритма, основанного на унитарных преобразованиях, сводится теперь к проблеме нахождения значений для a и $f(p, n)$.

Вещественные симметричные матрицы

18. Если A_0 — вещественная, то алгоритмы, которые мы будем использовать, дадут вещественные (и потому ортогональные) матрицы H_p . Если к тому же A_0 — симметричная, мы будем поддерживать точную симметрию всех $\bar{A}_p$ вычислением лишь их верхних треугольников и дополнением матрицы по симметрии. Заметим, что в этом случае матрицы F_p и Y_p в (16.3) точно симметричные, так как они представляют соответственно разности между $\bar{A}_p$ и $(H_p' + X_p)^T \bar{A}_{p-1} (H_p' + X_p)$ и $(H_p')^T \bar{A}_{p-1} H_p'$, а каждая из последних двух матриц точно симметричная. Соотношение (17.9) теперь

означает, что Z , которая точно симметричная, удовлетворяет неравенству

$$\|Z\|_2 \leq 2^{-t} \max |\lambda_i| \sum_{p=1}^s x_p. \quad (18.1)$$

Если $\lambda_1 + \delta\lambda_1, \lambda_2 + \delta\lambda_2, \dots, \lambda_n + \delta\lambda_n$ суть вычисленные собственные значения, то из гл. 2 § 44

$$\frac{|\delta\lambda_j|}{\max |\lambda_i|} \leq 2^{-t} \sum_{p=1}^s x_p, \quad (18.2)$$

так что наш анализ дает нам оценки для относительной ошибки в каждом λ_j в масштабе максимального по модулю собственного значения.

Мы повсюду дали анализ на основе 2-нормы; подобный анализ можно выполнить, используя евклидову норму. Тогда применение теоремы Виландта — Гофмана дает

$$[\sum (\delta\lambda_i)^2]^{1/2} \leq \|Z\|_E. \quad (18.3)$$

Мы хотели бы подчеркнуть, что анализ ошибок для несимметричных матриц, вообще говоря, не более труден, чем для симметричных. Он отличается лишь выводом оценки для Z , которая различна для этих случаев.

Ограничения унитарных преобразований

19. Рассуждения предыдущего параграфа могут создать впечатление, что алгоритм, основанный на унитарных преобразованиях, обязательно устойчив, но это далеко не так. Мы обсудили лишь те методы, в которых каждая матрица A_p получается одним элементарным преобразованием A_{p-1} , и предположили, что дополнительные нулевые элементы (если они есть), которые были введены в A_p , появились в результате этого одного преобразования. Мы показали, что устойчивость гарантирована при условии, что можно доказать на практике близость $\bar{H}_p$ и H'_p . Имелось лишь два источника ошибок:

(i) ошибки, сделанные при попытке вычислить элементы $\bar{A}_p$ из $\bar{H}_p^H \bar{A}_{p-1} \bar{H}_p$;

(ii) ошибки, соответствующие включению некоторых новых нулевых элементов в $\bar{A}_p$.

На практике всегда легко показать, что первая совокупность ошибок не вызывает затруднений. Что же касается второй совокупности, то здесь ситуация еще более благоприятная. Наш анализ был основан на отыскании малой оценки для $Y_p = \bar{A}_p - (H'_p) \bar{A}_{p-1} H'_p$.

Для определенных нулевых элементов соответствующие элементы Y_p точно равны нулю по определению H'_p . Однако этот практический путь проведения анализа не всегда дает наиболее точные оценки. Если мы можем доказать, что $\bar{H}_p$ близка к некоторой точно унитарной матрице R_p , можем получить хорошую оценку для $(R_p - \bar{H}_p)$ и, кроме того, можем показать, что элементы $(R_p^H \bar{A}_{p-1} R_p)$ в позициях, соответствующих нулевым элементам $\bar{A}_p$, малы, то мы можем выполнить подобный анализ, как в § 16, 17.

Однако существует несколько алгоритмов, основанных на элементарных унитарных преобразованиях, в которых нужные нулевые элементы

получаются лишь в окончательном произведении нескольких последовательных преобразований; появление этих нулей обычно зависит неявным образом от определения матриц H_p , содержащихся в алгоритме.

Предположим, что, начиная с A_0 , этим нулям следовало бы впервые появиться в матрице A_q после q преобразований. Если бы мы смогли доказать, что произведение $\bar{H}_1 \bar{H}_2 \dots \bar{H}_q$ близко к $H_1 H_2 \dots H_q$, то метод был бы устойчив. Однако если мы даже сможем показать, что каждая $\bar{H}_p$ близка к соответствующей H'_p (что мы обычно можем), то это еще не очень ценно, так как H'_p может быть сколь угодно далеко от H_p . Как мы заключили в § 14, матрицы A_p и $\bar{A}_p$ могут существенно отличаться друг от друга, и основным моментом нашего анализа было то, что не делается никакого сравнения между A_p и $\bar{A}_p$ или H_p и $\bar{H}_p$.

Если успех алгоритма решающим образом зависит от нашей возможности устанавливать нули в элементах $\bar{A}_q$, соответствующих нулевым элементам в A_q , то он может быть совсем бесполезен для практических целей. Мы дадим примеры катастрофической несостоятельности таких алгоритмов.

Анализ ошибок вычисления плоских вращений с плавающей запятой

20. Хотя существует значительное число методов, основанных на унитарных преобразованиях, имеется несколько основных вычислительных этапов, которые встречаются неоднократно, и в следующих небольших параграфах мы дадим анализ ошибок таких вычислений. Начнем с плоских вращений в стандартной арифметике с плавающей запятой, так как они, вероятно, самые простые. Мы ограничимся вещественным случаем; при этом рассматриваемые матрицы ортогональны.

В большинстве алгоритмов преобразование одного шага определяется следующим образом.

Имеем вектор $\bar{x}$ (обычно столбец или строка соответствующей преобразуемой матрицы) и хотим определить вращение R в плоскости (i, j) так, чтобы $(R\bar{x})_j = 0$.

Если мы напишем

$$\bar{x}^T = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n), \quad (20.1)$$

то компоненты $\bar{x}_r$ являются стандартными числами с плавающей запятой. Ссылаясь на гл. 1, § 43, мы видим, что точное вращение, определенное алгоритмом для вектора $\bar{x}$, таково, что

$$\left. \begin{aligned} r_{ii} &= \cos \theta, & r_{ij} &= \sin \theta, \\ r_{ji} &= -\sin \theta, & r_{jj} &= \cos \theta, \end{aligned} \right\} \quad (20.2)$$

где

$$\frac{\bar{x}_i}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}} = \cos \theta = c, \quad \frac{\bar{x}_j}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}} = \sin \theta = s.$$

Если $\bar{x}_j = 0$, мы берем $c = 1$, $s = 0$, независимо от того, равно ли $\bar{x}_i$ нулю или нет.

Назовем матрицу, определенную уравнениями (20.2), R , хотя в точном соответствии с предыдущими параграфами ее следовало бы назвать R' , так как это точная матрица, определенная алгоритмом для текущих значений вычисленных элементов. Но так как матрица, которая должна

была бы быть получена точным вычислением, никогда не входит в наше рассмотрение, мы опустим штрих в интересах более простого обозначения.

На практике мы получаем матрицу $\bar{R}$ с элементами $\bar{c}$ и $\bar{s}$ на месте c и s , где

$$\bar{c} = fl\left(\frac{\bar{x}_i}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}}\right), \quad \bar{s} = fl\left(\frac{\bar{x}_j}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}}\right). \quad (20.3)$$

Вычисление $\bar{c}$ и $\bar{s}$ осуществляется по шагам:

$$a = fl(\bar{x}_i^2 + \bar{x}_j^2), \quad b = fl(a^{1/2}), \quad c = fl\left(\frac{\bar{x}_i}{b}\right), \quad s = fl\left(\frac{\bar{x}_j}{b}\right). \quad (20.4)$$

Если $\bar{x}_j = 0$, мы берем $\bar{c} = 1$, $\bar{s} = 0$, так что в этом случае $\bar{R} = R$.

Для того чтобы проиллюстрировать применение нашего метода анализа плавающей запятой, рассмотрим (20.4) подробно. В дальнейших примерах, которые так же просты, как этот, дадим лишь окончательный результат. Для того чтобы получить как можно более точную оценку, вернемся к оценке вида

$$(1 - 2^{-t})^r \leq 1 + \varepsilon \leq (1 + 2^{-t})^r, \quad (20.5)$$

вместо того чтобы использовать более простую оценку вида 2^{-t} . Имеем

$$(i) \quad a = fl(\bar{x}_i^2 + \bar{x}_j^2) \equiv \bar{x}_i^2(1 + \varepsilon_1) + \bar{x}_j^2(1 + \varepsilon_2) \equiv (\bar{x}_i^2 + \bar{x}_j^2)(1 + \varepsilon_3),$$

где

$$(1 - 2^{-t})^2 \leq 1 + \varepsilon_i \leq (1 + 2^{-t})^2 \quad (i = 1, 2, 3).$$

$$(ii) \quad b = fl(a^{1/2}) \equiv a^{1/2}(1 + \varepsilon_4),$$

где $|\varepsilon_4| < (1,00001) 2^{-t}$ из (10.2).

$$(iii) \quad \bar{c} = fl\left(\frac{\bar{x}_i}{b}\right) \equiv \bar{x}_i \frac{1 + \varepsilon_5}{b}, \quad \bar{s} = fl\left(\frac{\bar{x}_j}{b}\right) \equiv \bar{x}_j \frac{1 + \varepsilon_6}{b},$$

где

$$|\varepsilon_i| \leq 2^{-t} \quad (i = 5, 6).$$

Комбинируя (i), (ii) и (iii), получаем

$$\bar{c} \equiv c \frac{1 + \varepsilon_5}{(1 + \varepsilon_3)^{1/2}(1 + \varepsilon_4)}, \quad \bar{s} \equiv s \frac{1 + \varepsilon_6}{(1 + \varepsilon_3)^{1/2}(1 + \varepsilon_4)}. \quad (20.6)$$

Простейшие вычисления показывают, что

$$\bar{c} \equiv c(1 + \varepsilon_7), \quad \bar{s} \equiv s(1 + \varepsilon_8), \quad (20.7)$$

$$|\varepsilon_i| < (3,003) 2^{-t} \quad (i = 7, 8), \quad (20.8)$$

где множитель 3,003 максимально возможный. Вычисленные $\bar{c}$ и $\bar{s}$, следовательно, имеют небольшую относительную ошибку. Должно быть подчеркнуто однако, что использование вычислений с плавающей запятой не всегда гарантирует результаты, имеющие малую относительную ошибку, даже когда вычисления так же просты, как те, которые мы только что анализировали (гл. 5, § 12).

Мы можем написать

$$\bar{c} = c + \delta c, \quad \bar{s} = s + \delta s, \quad \bar{R} = R + \delta R. \quad (20.9)$$

Здесь δR — нулевая матрица, за исключением пересечений i -й и j -й строк и столбцов, где она имеет элементы

$$\begin{bmatrix} \delta c & \delta s \\ -\delta s & \delta c \end{bmatrix}. \quad (20.10)$$

Следовательно, имеем

$$\|\delta R\|_2 = (\delta c^2 + \delta s^2)^{1/2} < (3,003) 2^{-t}, \quad (20.11)$$

$$\|\delta R\|_E = [2(\delta c^2 + \delta s^2)]^{1/2} < (3,003) 2^{1/2} 2^{-t}, \quad (20.12)$$

показывая, что вычисленная $\bar{R}$ близка к точной R , соответствующей данным $\bar{x}_i$ и $\bar{x}_j$. Если даны два числа $\bar{x}_i$ и $\bar{x}_j$ с плавающей запятой, мы назовем $\bar{R}$ соответствующим приближенным вращением в плоскости (i, j) .

Умножение на плоское вращение

21. Рассмотрим теперь сложную операцию, которую определим следующим образом. Даны два числа $\bar{x}_i$ и $\bar{x}_j$ с плавающей запятой и матрица A , имеющая стандартные элементы также с плавающей запятой; вычисляем приближенное вращение $\bar{R}$ в плоскости (i, j) и затем вычисляем $fl(\bar{R}A)$.

Для простоты записи рассмотрим сначала вычисление $fl(\bar{R}a)$, где a — вектор. Изменяются лишь те два элемента a , которые находятся в позициях i и j .

Если мы напомним

$$b = fl(\bar{R}a), \quad (21.1)$$

то

$$\left. \begin{aligned} b_p &= a_p \quad (p \neq i, j), \\ b_i &= fl(\bar{c}a_i + \bar{s}a_j) \equiv \bar{c}a_i(1 + \varepsilon_1) + \bar{s}a_j(1 + \varepsilon_2), \\ b_j &= fl(-\bar{s}a_i + \bar{c}a_j) \equiv -\bar{s}a_i(1 + \varepsilon_3) + \bar{c}a_j(1 + \varepsilon_4), \\ (1 - 2^{-t})^2 &\leq 1 + \varepsilon_i \leq (1 + 2^{-t})^2 \quad (i = 1, 2, 3, 4). \end{aligned} \right\} \quad (21.2)$$

Мы можем выразить это в виде

$$\begin{bmatrix} b_i \\ b_j \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \cdot \begin{bmatrix} a_i \\ a_j \end{bmatrix} + \begin{bmatrix} \delta c & \delta s \\ -\delta s & \delta c \end{bmatrix} \cdot \begin{bmatrix} a_i \\ a_j \end{bmatrix} + \begin{bmatrix} \bar{c}a_i\varepsilon_1 + \bar{s}a_j\varepsilon_2 \\ -\bar{s}a_i\varepsilon_3 + \bar{c}a_j\varepsilon_4 \end{bmatrix}. \quad (21.3)$$

Используя оценки для δc , δs и ε_i и записывая

$$fl(\bar{R}a) - Ra \equiv f, \quad (21.4)$$

имеем

$$\begin{aligned} \|f\|_2 &\leq (\delta c^2 + \delta s^2)^{1/2} (a_i^2 + a_j^2)^{1/2} + (2,0002) 2^{-t} [2(\bar{c}^2 + \bar{s}^2)]^{1/2} (a_i^2 + a_j^2)^{1/2} \leq \\ &\leq 6 \cdot 2^{-t} (a_i^2 + a_j^2)^{1/2}, \end{aligned} \quad (21.5)$$

и опять множитель 6 максимально возможный. Заметим, что это означает, что если a_i и a_j очень малы по отношению к некоторым другим компонентам, то $\|f\|_2/\|a\|_2$ очень мало. Мы найдем, что это неверно для вычислений с фиксированной запятой. Из (21.3) и соотношения $\|x + y + z\|_2 \leq \|x\|_2 + \|y\|_2 + \|z\|_2$ получаем

$$(b_i^2 + b_j^2)^{1/2} \leq (a_i^2 + a_j^2)^{1/2} + 6 \cdot 2^{-t} (a_i^2 + a_j^2)^{1/2} = (1 + 6 \cdot 2^{-t}) (a_i^2 + a_j^2)^{1/2}. \quad (21.6)$$

Это соотношение важно для последующего анализа.

Рассмотрение (20.11) и (21.5) показывает, что половина ошибки происходит от ошибок, сделанных при вычислении $\bar{R}$, а другая половина — от ошибок, сделанных при умножении на вычисленную $\bar{R}$.

Если вектор a есть вектор $\bar{x}$, из которого были получены элементы $\bar{x}_i$ и $\bar{x}_j$, то на практике обычно не вычисляют $fl(\bar{R}x)$ и вместо этого i -ю компоненту берут равной $fl[(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}]$, а j -ю компоненту — равной нулю. Выражение, определяющее i -ю компоненту, будет уже вычислено при нахождении $\bar{R}$.

Два элемента f поэтому суть

$$fl[(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}] - (\bar{x}_i^2 + \bar{x}_j^2)^{1/2}$$

и нуль, и легко может быть проверено, что оценка (21.5) вполне пригодна в этом случае. Поэтому не стоит беспокоиться о специальном рассмотрении вектора $\bar{x}$.

Наш результат немедленно распространяется на вычисление $fl(\bar{R}A)$, где A есть $(n \times m)$ -матрица. Все m столбцов $\bar{R}A$ преобразуются независимо, и мы имеем

$$fl(\bar{R}A) - RA \equiv F, \quad (21.7)$$

где F — нулевая матрица, за исключением i -й и j -й строк, и

$$\|F\|_E \leq 6 \cdot 2^{-t} (a_{i1}^2 + a_{j1}^2 + a_{i2}^2 + \dots + a_{im}^2 + a_{jm}^2)^{1/2}. \quad (21.8)$$

Заметим, что это оценка для $\|F\|_E$, а не для $\|F\|_2$. Квадраты норм i -й и j -й строк заменяют a_i^2 и a_j^2 в нашем анализе в случае вектора. Оценка (21.8) пригодна даже тогда, когда один из столбцов A является нашим вектором $\bar{x}$ и преобразуется специально.

Умножение на последовательность плоских вращений

22. Предположим, что нам дана последовательность s пар чисел с плавающей запятой $\bar{x}_{ip}$, $\bar{x}_{jp}$ ($p = 1, \dots, s$) и матрица A_0 . Соответственно каждой паре мы можем вычислить приближенное плоское вращение $\bar{R}_p$ в плоскости (i_p, j_p) и затем вычислить s матриц, определенных равенствами

$$\left. \begin{aligned} \bar{A}_1 &= fl(\bar{R}_1 A_0) \equiv R_1 A_0 + F_1, \\ \bar{A}_p &= fl(\bar{R}_p \bar{A}_{p-1}) \equiv R_p \bar{A}_{p-1} + F_p \quad (p = 2, \dots, s). \end{aligned} \right\} \quad (22.1)$$

Комбинируя равенства (22.1) для последовательности значений p , получим

$$\bar{A}_p \equiv Q_1 A_0 + Q_2 F_1 + Q_3 F_2 + \dots + Q_p F_{p-1} + F_p, \quad (22.2)$$

где

$$Q_k = R_p R_{p-1} \dots R_k.$$

Так как по определению R_k точно ортогональны, то Q_k также точно ортогональны. Следовательно,

$$\left\| \bar{A}_p - R_p R_{p-1} \dots R_1 A_0 \right\| \leq \|F_1\| + \|F_2\| + \dots + \|F_p\| \quad (p = 1, \dots, s) \quad (22.3)$$

для 2-нормы или евклидовой нормы.

На практике будем часто иметь дело с множеством $\frac{1}{2} n (n - 1)$ вращений в последовательных плоскостях $(1, 2) (1, 3) \dots (1, n); (2, 3) (2, 4) \dots (2, n); \dots; (n - 1, n)$.

Покажем, что для специального множества вращений этого вида можно получить очень удовлетворительную оценку для

$$\| \bar{A}_N - R_N R_{N-1} \dots R_1 A_0 \|_E \quad \left(N = \frac{1}{2} n (n - 1) \right). \quad (22.4)$$

23. Основная трудность, связанная с анализом, является трудностью записи, поэтому начнем сначала с отдельного вектора вместо матрицы A_0 и для большей простоты возьмем $n = 5$ (и, следовательно, $N = 10$); доказательство в общем случае достаточно очевидно вытекает отсюда. Обозначим начальный вектор через v_0 , а последовательно полученные векторы через $v_1, v_2, \dots, v_N$. В табл. 1 мы даем элементы векторов $\bar{v}_i$ и, для того

Т а б л и ц а 1										
v_0	$\bar{v}_1$	$\bar{v}_2$	$\bar{v}_3$	$\bar{v}_4$	$\bar{v}_5$	$\bar{v}_6$	$\bar{v}_7$	$\bar{v}_8$	$\bar{v}_9$	$\bar{v}_{10}$
a_0	$\bar{a}_1$	$\bar{a}_2$	$\bar{a}_3$	$\bar{a}_4$	$\bar{a}_4$	$\bar{a}_4$	$\bar{a}_4$	$\bar{a}_4$	$\bar{a}_4$	$\bar{a}_4$
b_0	$\bar{b}_1$	$\bar{b}_1$	$\bar{b}_1$	$\bar{b}_1$	$\bar{b}_2$	$\bar{b}_3$	$\bar{b}_4$	$\bar{b}_4$	$\bar{b}_4$	$\bar{b}_4$
c_0	$\bar{c}_0$	$\bar{c}_1$	$\bar{c}_1$	$\bar{c}_1$	$\bar{c}_2$	$\bar{c}_2$	$\bar{c}_2$	$\bar{c}_3$	$\bar{c}_4$	$\bar{c}_4$
d_0	$\bar{d}_0$	$\bar{d}_0$	$\bar{d}_1$	$\bar{d}_1$	$\bar{d}_1$	$\bar{d}_2$	$\bar{d}_2$	$\bar{d}_3$	$\bar{d}_3$	$\bar{d}_4$
e_0	$\bar{e}_0$	$\bar{e}_0$	$\bar{e}_0$	$\bar{e}_1$	$\bar{e}_1$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$

чтобы избежать путаницы с индексами, используем различные буквы для элементов в каждой позиции. Мы изменяем индекс только тогда, когда изменяется соответствующий элемент. В общем случае первый элемент изменяется в каждом из $(n - 1)$ первых преобразований и затем последующим изменениям не подвергается. Аналогично второй элемент изменяется в каждом из $(n - 2)$ следующих преобразований и последующим изменениям не подвергается и т. д. Мы указали эти группы в табл. 1. Каждый элемент всегда изменяется $(n - 1)$ раз, так что окончательный индекс будет $(n - 1)$ для каждого элемента.

Если мы запишем

$$\bar{v}_p = fl(\bar{R}_p \bar{v}_{p-1}) \equiv R_p \bar{v}_{p-1} + f_p, \quad (23.1)$$

то предыдущий анализ показывает, что

$$\| \bar{v}_N - R_N R_{N-1} \dots R_1 v_0 \|_2 \leq \| f_1 \|_2 + \dots + \| f_N \|_2. \quad (23.2)$$

Теперь анализ § 21 немедленно применяется к каждому преобразованию, так как каждый $\bar{v}_p$ есть вектор с компонентами с плавающей запятой, и, полагая $6 \cdot 2^{-t} = x$, мы имеем

$$\left. \begin{aligned} \|f_1\| &\leq x(a_0^2 + b_0^2)^{1/2}, & \|f_2\| &\leq x(\bar{a}_1^2 + \bar{c}_0^2)^{1/2}, & \|f_3\| &\leq x(\bar{a}_2^2 + \bar{d}_0^2)^{1/2}, & \|f_4\| &\leq x(\bar{a}_3^2 + \bar{e}_0^2)^{1/2}, \\ \|f_5\| &\leq x(\bar{b}_1^2 + \bar{c}_1^2)^{1/2}, & \|f_6\| &\leq x(\bar{b}_2^2 + \bar{d}_1^2)^{1/2}, & \|f_7\| &\leq x(\bar{b}_3^2 + \bar{e}_1^2)^{1/2}, \\ & & \|f_8\| &\leq x(\bar{c}_2^2 + \bar{d}_2^2)^{1/2}, & \|f_9\| &\leq x(\bar{c}_3^2 + \bar{e}_2^2)^{1/2}, \\ & & & & \|f_{10}\| &\leq x(\bar{d}_3^2 + \bar{e}_3^2)^{1/2}, \end{aligned} \right\} \quad (23.3)$$

где группировка выполнена для того, чтобы подчеркнуть общий вид. Если мы запишем

$$f = \|f_1\|_2 + \dots + \|f_N\|_2 = xS, \quad (23.4)$$

то, применяя неравенство

$$(q_1 + q_2 + \dots + q_N)^2 \leq N(q_1^2 + q_2^2 + \dots + q_N^2), \quad (23.5)$$

которое справедливо для всех положительных q_i , имеем для (23.3)

$$S^2 \leq 10 \left[\begin{array}{c} (a_0^2 + b_0^2) + (\bar{a}_1^2 + \bar{c}_0^2) + (\bar{a}_2^2 + \bar{d}_0^2) + (\bar{a}_3^2 + \bar{e}_0^2) + \\ + (\bar{b}_1^2 + \bar{c}_1^2) + (\bar{b}_2^2 + \bar{d}_1^2) + (\bar{b}_3^2 + \bar{e}_1^2) + \\ + (\bar{c}_2^2 + \bar{d}_2^2) + (\bar{c}_3^2 + \bar{e}_2^2) + \\ + (\bar{d}_3^2 + \bar{e}_3^2) \end{array} \right], \quad (23.6)$$

где множитель 10 есть N в общем случае.

Записывая для краткости $k = (1+x)^2$, соотношения (21.6), соответствующие для каждого из N преобразований, будут таковы:

$$\left. \begin{array}{l} \bar{a}_1^2 + \bar{b}_1^2 \leq k(a_0^2 + b_0^2), \\ \bar{a}_2^2 + \bar{c}_1^2 \leq k(\bar{a}_1^2 + \bar{c}_0^2), \quad \bar{b}_2^2 + \bar{c}_2^2 \leq k(\bar{b}_1^2 + \bar{c}_1^2), \quad \bar{c}_3^2 + \bar{d}_3^2 \leq k(\bar{c}_2^2 + \bar{d}_2^2), \\ \bar{a}_3^2 + \bar{d}_1^2 \leq k(\bar{a}_2^2 + \bar{d}_0^2), \quad \bar{b}_3^2 + \bar{d}_2^2 \leq k(\bar{b}_2^2 + \bar{d}_1^2), \quad \bar{c}_4^2 + \bar{e}_3^2 \leq k(\bar{c}_3^2 + \bar{e}_2^2), \\ \bar{a}_4^2 + \bar{e}_1^2 \leq k(\bar{a}_3^2 + \bar{e}_0^2), \quad \bar{b}_4^2 + \bar{e}_2^2 \leq k(\bar{b}_3^2 + \bar{e}_1^2), \quad \bar{d}_4^2 + \bar{e}_4^2 \leq k(\bar{d}_3^2 + \bar{e}_3^2). \end{array} \right\} \quad (23.7)$$

Мы хотим установить соотношение между S и $\|v_0\|_2$. Из соотношений в первом столбце (23.7) имеем

$$\left. \begin{array}{l} \bar{a}_1^2 \leq k(a_0^2 + b_0^2) \quad - \bar{b}_1^2, \\ \bar{a}_2^2 \leq k^2(a_0^2 + b_0^2) + k\bar{c}_0^2 \quad - k\bar{b}_1^2 - \bar{c}_1^2, \\ \bar{a}_3^2 \leq k^3(a_0^2 + b_0^2) + k^2\bar{c}_0^2 + k\bar{d}_0^2 - k^2\bar{b}_1^2 - k\bar{c}_1^2 - \bar{d}_1^2 \end{array} \right\} \quad (23.8)$$

и, следовательно, определяя S_r равенством

$$S_r = 1 + k + k^2 + \dots + k^{r-1}, \quad (23.9)$$

видим, что сумма в первой строке правой части (23.6) будет ограничена величиной

$$S_4 a_0^2 + S_4 b_0^2 + S_3 \bar{c}_0^2 + S_2 \bar{d}_0^2 + S_1 \bar{e}_0^2 - S_3 \bar{b}_1^2 - S_2 \bar{c}_1^2 - S_1 \bar{d}_1^2. \quad (23.10)$$

Точно так же оставшиеся три суммы в (23.6) ограничены величинами

$$\left. \begin{array}{l} S_3 \bar{b}_1^2 + S_3 \bar{c}_1^2 + S_2 \bar{d}_1^2 + S_1 \bar{e}_1^2 - S_2 \bar{c}_2^2 - S_1 \bar{d}_2^2, \\ S_2 \bar{c}_2^2 + S_2 \bar{d}_2^2 + S_1 \bar{e}_2^2 \quad - S_1 \bar{d}_3^2, \\ S_1 \bar{d}_3^2 + S_1 \bar{e}_3^2 \end{array} \right\} \quad (23.11)$$

где наша группировка указывает соответствующий результат для общего случая. Суммируя выражения в (23.10) и (23.11), мы видим, что (23.6) эквивалентно

$$S^2 \leq 10 \left[\begin{array}{c} S_4 a_0^2 + S_4 b_0^2 + S_3 \bar{c}_0^2 + S_2 \bar{d}_0^2 + S_1 \bar{e}_0^2 + \\ + k^2 \bar{c}_1^2 + k \bar{d}_1^2 + \bar{e}_1^2 + \\ + k \bar{d}_2^2 + \bar{e}_2^2 + \\ + \bar{e}_3^2 \end{array} \right], \quad (23.12)$$

так как имеет место значительное сокращение между отрицательными членами каждой строки и положительными членами последующей строки. Первая строка целиком выражена в значениях элементов v_0 . Чтобы привести вторую строку к аналогичному виду, умножим неравенства в первом столбце (23.7) соответственно на k^3 , k^2 , k , 1 и сложим. Это дает

$$\bar{a}_4^2 + k^3 \bar{b}_1^2 + k^2 \bar{c}_1^2 + k \bar{d}_1^2 + \bar{e}_1^2 \leq k^4 (a_0^2 + b_0^2) + k^3 c_0^2 + k^2 d_0^2 + k e_0^2 \quad (23.13)$$

и, следовательно, тем более

$$k^3 \bar{b}_1^2 + (k^2 \bar{c}_1^2 + k \bar{d}_1^2 + \bar{e}_1^2) \leq k^4 (a_0^2 + b_0^2) + k^3 c_0^2 + k^2 d_0^2 + k e_0^2. \quad (23.14)$$

Аналогично для второго столбца (23.7)

$$\begin{aligned} k^2 \bar{c}_2^2 + (k \bar{d}_2^2 + \bar{e}_2^2) &\leq k^3 (\bar{b}_1^2 + \bar{c}_1^2) + k^2 \bar{d}_1^2 + k e_1^2 \leq \\ &\leq k^5 (a_0^2 + b_0^2) + k^4 c_0^2 + k^3 d_0^2 + k^2 e_0^2, \end{aligned} \quad (23.15)$$

где последняя строка заведомо следует из умножения (23.14) на k . Аналогично, используя (23.15) для третьего столбца (23.7), получим

$$k \bar{d}_3^2 + (\bar{e}_3^2) \leq k^2 (\bar{c}_2^2 + \bar{d}_2^2) + k e_2^2 \leq k^6 (a_0^2 + b_0^2) + k^5 c_0^2 + k^4 d_0^2 + k^3 e_0^2. \quad (23.16)$$

Выражения в скобках левых частей (23.14), (23.15) и (23.16) являются второй, третьей и четвертой строками (23.12), и, следовательно,

$$S^2 \leq 10(S_7 a_0^2 + S_7 b_0^2 + S_6 c_0^2 + S_5 d_0^2 + S_4 e_0^2), \quad (23.17)$$

а из способа доказательства ясно, что в общем случае коэффициенты будут S_{2n-3} , S_{2n-3} , S_{2n-4} , ..., S_{n-1} . Заменяя каждый из коэффициентов в (23.17) на максимальный S_7 , имеем

$$S^2 \leq 10 S_7 (a_0^2 + b_0^2 + c_0^2 + d_0^2 + e_0^2) \quad (23.18)$$

и, следовательно, в общем случае

$$\begin{aligned} S^2 &\leq \frac{1}{2} n(n-1) S_{2n-3} \|v_0\|_2^2 \leq \frac{1}{2} n(n-1)(2n-3)(1+x)^{4n-8} \|v_0\|_2^2 \leq \\ &\leq n^3 (1+x)^{4n-8} \|v_0\|_2^2, \end{aligned} \quad (23.19)$$

так как

$$S_r \leq r(1+x)^{2r-2}.$$

Ссылаясь на (23.2) и (23.4), мы видим окончательно, что отсюда следует неравенство

$$\|\bar{v}_N - R_N R_{N-1} \dots R_1 v_0\|_2 \leq x n^{3/2} (1+x)^{2n-4} \|v_0\|_2. \quad (23.20)$$

24. Анализ немедленно распространяется на левое умножение $(n \times m)$ -матрицы на те же N приближенных вращений. Мы должны только заменить квадрат i -го элемента обычного вектора на сумму квадратов элементов в i -й строке текущей преобразованной матрицы на каждом этапе доказательства. Выражение, соответствующее (23.6), все равно имеет лишь $\frac{1}{2} n(n-1)$ членов в круглых скобках внутри больших квадратных скобок. В этом случае имеем

$$\begin{aligned} \|\bar{A}_N - R_N R_{N-1} \dots R_1 A_0\|_E &\leq x n^{3/2} (1+x)^{2n-4} \|A_0\|_E = \\ &= 6 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{2n-4} \|A_0\|_E. \end{aligned} \quad (24.1)$$

Множитель $(1 + 6 \cdot 2^{-t})^{2n-4}$ экспоненциально стремится к бесконечности вместе с n , но, к счастью, это неважно, так как вычисленная $\bar{A}_N$ будет иметь ценность лишь до тех пор, пока отношение $\| \text{ошибка} \|_E / \| A_0 \|_E$ мало. Предположим, что мы хотим чтобы это отношение было меньше 10^{-6} . Тогда мы должны иметь

$$6 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{2n-4} \leq 10^{-6}. \quad (24.2)$$

Если мы считаем t фиксированным, то неравенство дает самое большое значение n , для которого можно гарантировать приемлемый результат. Очевидно, что мы должны иметь $6 \cdot 2^{-t} n^{3/2} < 10^{-6}$, и, следовательно,

$$(1 + 6 \cdot 2^{-t})^{2n-4} < (1 + 10^{-6} n^{-3/2})^{2n} < \exp(2 \cdot 10^{-6} n^{-1/2}) < (1 + 2 \cdot 10^{-6}). \quad (24.3)$$

Поэтому этот множитель совсем незначителен в своем влиянии на предел применимости.

Правая часть (24.1) представляет собой максимальную верхнюю оценку, и существует несколько моментов в анализе, когда были использованы неравенства, очень слабые для некоторых распределений элементов матрицы. Можно ожидать, что статистическое распределение ошибок округления существенно понизит уровень нашей оценки, и только по этой причине множитель порядка $n^{3/4}$ вместо $n^{3/2}$, может быть, будет более реален для больших n .

Заметим, что при получении этой оценки мы не предполагали никакой связи между $\bar{x}_i$ и $\bar{x}_j$, определяющими вращения, и матрицами A_0 и $\bar{A}_p$. Наш результат справедлив при любом происхождении этой пары чисел с плавающей запятой. Каждая пара могла быть получена из текущей $\bar{A}_p$ или же из совершенно независимого источника. Тем не менее наше сравнение делается между точным преобразованием, соответствующим данным парам чисел, и вычисленным преобразованием. Если каждая пара берется из текущей $\bar{A}_p$, то не делается никакого сравнения между окончательно вычисленной $\bar{A}_N$ и матрицей, которая могла бы быть получена, если бы всюду выполнялись точные вычисления, *включая выбор пар чисел из точно преобразованных матриц*.

Ошибка в произведении приближенных плоских вращений

25. До сих пор мы рассматривали лишь левые умножения и неподобные преобразования. Прежде чем распространить наш результат, заметим, что эти односторонние преобразования имеют и самостоятельный интерес. В следующей главе мы рассмотрим приведение квадратной матрицы к треугольной форме левыми умножениями на элементарные ортогональные матрицы. Результат § 24 немедленно применяется к этим вычислениям.

В проблеме собственных векторов нам будет нужно вычислять произведение последовательности приближенных плоских вращений и будет интересно оценить отклонение вычисленного произведения от ортогональной матрицы. Беря $A_0 = I$ в (24.1), имеем

$$\| \bar{P}_N - R_N R_{N-1} \dots R_1 \|_E \leq x n^{3/2} (1 + x)^{2n-4} \| I \|_E = x n^2 (1 + x)^{2n-4}, \quad (25.1)$$

где $\bar{P}_N$ — вычисленное произведение; матрица $R_N R_{N-1} \dots R_1$ точно ортогональная. Если мы обозначим i -е столбцы $\bar{P}_N$ и $R_N R_{N-1} \dots R_1$

соответственно через $\bar{p}_i$ и p_i , то, так как каждый столбец матрицы I преобразуется независимо, применение (23.20) показывает, что

$$\|\bar{p}_i - p_i\|_2 \leq xn^{3/2}(1+x)^{2n-4} \|e_i\|_2 = xn^{3/2}(1+x)^{2n-4} \quad (25.2)$$

и, следовательно,

$$1 - xn^{3/2}(1+x)^{2n-4} \leq \|\bar{p}_i\|_2 \leq 1 + xn^{3/2}(1+x)^{2n-4}. \quad (25.3)$$

Итак, если мы напомним

$$\bar{p}_i = p_i + q_i, \quad (24.4)$$

то

$$\bar{p}_j^T \bar{p}_i = (p_j^T + q_j^T)(p_i + q_i) = q_j^T p_i + p_j^T q_i + q_j^T q_i, \quad (25.5)$$

$$\begin{aligned} |\bar{p}_j^T \bar{p}_i| &\leq \|q_j\|_2 + \|q_i\|_2 + \|q_j\|_2 \|q_i\|_2 \leq \\ &\leq xn^{3/2}(1+x)^{2n-4} [2 + xn^{3/2}(1+x)^{2n-4}]. \end{aligned} \quad (25.6)$$

Это неравенство дает оценку отклонения столбца P_N от столбца ортогональной матрицы. Например, если $t = 40$ и $n = 100$, мы имеем

$$xn^{3/2} = 6 \cdot 2^{-40} \cdot 1000 < 6 \cdot 10^{-9},$$

и, следовательно, вычисленное произведение 100 приближенных вращений будет, конечно, ортогональной матрицей с точностью восемь десятичных знаков. Статистические соображения показывают, что оно будет ортогональной матрицей с точностью около десяти знаков.

Ошибки в подобных преобразованиях

26. Если B_0 есть $(m \times n)$ -матрица, то очевидно, что существует точно такой же анализ ошибок для правого умножения B_0 на совокупность N транспонированных приближенных вращений. Поэтому, если $\bar{B}_N$ — окончательная матрица, то, как и в (24.1), имеем

$$\|\bar{B}_N - B_0 R_1^T R_2^T \dots R_N^T\|_E \leq xn^{3/2}(1+x)^{2n-4} \|B_0\|_E. \quad (26.1)$$

Теперь предположим, что мы начинаем с A_0 и выполняем N левых умножений, чтобы получить $\bar{A}_N$. Эта матрица имеет элементы с плавающей запятой и, следовательно, может быть использована как B_0 в (26.1); при этом

$$\|\bar{B}_N - \bar{A}_N R_1^T R_2^T \dots R_N^T\|_E \leq xn^{3/2}(1+x)^{2n-4} \|\bar{A}_N\|_E = y \|\bar{A}_N\|_E. \quad (26.2)$$

Кроме того, из (24.1) мы имеем в той же самой сокращенной записи

$$\|\bar{A}_N - R_N R_{N-1} \dots R_1 A_0\|_E \leq y \|A_0\|_E \quad (26.3)$$

и, следовательно,

$$\|\bar{A}_N\|_E \leq (1+y) \|A_0\|_E. \quad (26.4)$$

Объединяя эти результаты, мы получаем

$$\begin{aligned} \|\bar{B}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_E &\leq \|\bar{B}_N - \bar{A}_N R_1^T \dots R_N^T\|_E + \\ &+ \|(\bar{A}_N - R_N \dots R_1 A_0) R_1^T \dots R_N^T\|_E \leq \\ &\leq y \|\bar{A}_N\|_E + \|\bar{A}_N - R_N \dots R_1 A_0\|_E \leq \\ &\leq y(1+y) \|A_0\|_E + y \|A_0\|_E = \\ &= y(2+y) \|A_0\|_E, \end{aligned} \quad (26.5)$$

показывая, что окончательная матрица отличается от точного подобного преобразования A_0 на матрицу с нормой, ограниченной величиной

$$6 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{2n-4} [2 + 6 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{2n-4}] \|A_0\|_E. \quad (26.6)$$

Так как вычисленная матрица $\bar{B}_N$ будет интересна только тогда, когда y мало, то наличие члена y в скобках (26.5) неважно, и относительная ошибка лишь вдвое больше, чем при односторонних преобразованиях, как мы и могли ожидать.

27. Существует несколько алгоритмов, в которых левые умножения заканчиваются прежде, чем выполнится какое-нибудь из правосторонних преобразований, но часто это невозможно, так как p -я пара чисел, которая определяет p -е вращение, берется из матрицы, полученной выполнением первых $(p - 1)$ подобных преобразований.

Начиная с A_0 , последовательность может быть определена так:

$$\left. \begin{aligned} \bar{X}_1 &= fl(\bar{R}_1 A_0), & \bar{A}_1 &= fl(\bar{X}_1 \bar{R}_1^T), \\ \bar{X}_p &= fl(\bar{R}_p \bar{A}_{p-1}), & \bar{A}_p &= fl(\bar{X}_p \bar{R}_p^T) \quad (p = 2, \dots, N). \end{aligned} \right\} \quad (27.1)$$

На первый взгляд кажется возможным, что выполнение правого умножения между двумя левыми могло нарушить прежнее особое соотношение между нормами строк $\bar{A}_p$, которое было существенно при получении оценки. Однако рассмотрим эффект включения точного правого умножения на точное вращение между двумя левыми умножениями. Правое умножение действует на два столбца, оставляя норму каждой строки неизменной; далее, вклад ошибки, сделанной при любом левом умножении, зависит лишь от норм двух строк, которые оно изменяет, и, следовательно, включение любого числа точных правых умножений не оказывает никакого влияния на оценку, полученную для ошибки. Это показывает, что порядок выполнения умножений дает эффект лишь второго порядка. Доказательство, аналогичное доказательству в § 23, показывает, что

$$\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_E \leq 12 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{4n-7} \|A_0\|_E. \quad (27.2)$$

Что касается множителя $(1 + 6 \cdot 2^{-t})^{4n-7}$, то его оценка несколько слаба и, возможно, может быть улучшена, но так как вычисление имеет ценность лишь тогда, когда $12 \cdot 2^{-t} n^{3/2}$ существенно меньше единицы, то данный множитель никогда не может иметь практической важности.

Этот окончательный результат оправдывает для плоских вращений наше замечание в § 17, что ошибки округления, сделанные в действительных умножениях, не приводят к численной неустойчивости. Мы дали анализ большой общности, чтобы охватить широкий ряд возможных приложений. В большинстве алгоритмов последовательные преобразования будут иметь все больше и больше нулевых элементов, и, следовательно, для всякого специального приложения обычно может быть получена несколько лучшая оценка. Однако такие улучшения, как правило, делают более эффективными не основной член $n^{3/2} 2^{-t}$ в оценке, а лишь постоянный множитель.

Симметричные матрицы

28. Если исходная матрица симметричная, мы будем брать все $\bar{A}_p$ точно симметричными. Поэтому наш анализ не сразу применим, так как если вращение выполняется в плоскости (i, j) , то элементы ниже диагонали не вычисляются в действительности при каждом преобразовании, а им

автоматически приписывается то же самое значение, что и элементам выше диагонали. Если мы предполагаем, что $\bar{A}_{p-1}$ симметричная, то даже если вычисления в действительности были выполнены с использованием приближенного плоского вращения, точная симметрия должна будет сохраниться, за исключением, возможно, элементов в позициях (i, j) и (j, i) .

По-видимому, невозможно показать каким-либо простым способом, что оценка ошибки, которую мы получили для несимметричного случая, пригодна для распространения на симметричный случай, но сравнительно просто показать, что она увеличится не более чем в $2^{1/2}$. Основной момент состоит в следующем.

Если мы выполним полное преобразование в плоскости (1,2), используя сначала или левое или правое умножение, то матрица ошибок будет нулевая, за исключением 1-й и 2-й строк и столбцов, и будет симметричная, кроме элементов, стоящих на пересечении этих строк и столбцов. Мы можем обозначить ошибки в этих позициях через

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} p & r \\ q & s \end{bmatrix} \quad (28.1)$$

соответственно. Обе матрицы ошибок будут одинаковыми в других местах этих строк и столбцов. Теперь, если мы предполагаем полную симметрию ошибок в позициях, соответствующих (28.1), то будет

$$\begin{bmatrix} p & q \\ q & s \end{bmatrix} \quad \text{или} \quad \begin{bmatrix} p & r \\ r & s \end{bmatrix}. \quad (28.2)$$

Сумма квадратов ошибок для каждого из четырех случаев есть соответственно

$$\left. \begin{array}{l} \text{(i)} \quad p^2 + q^2 + r^2 + s^2 + \Sigma^2 = k_1 \\ \text{(ii)} \quad p^2 + q^2 + r^2 + s^2 + \Sigma^2 = k_1 \\ \text{(iii)} \quad p^2 + q^2 + q^2 + s^2 + \Sigma^2 = k_3 \\ \text{(iv)} \quad p^2 + r^2 + r^2 + s^2 + \Sigma^2 = k_4 \end{array} \right\}, \quad (28.3)$$

где Σ^2 — сумма квадратов остальных ошибок (она одинакова во всех четырех случаях). Поэтому мы окончательно имеем

$$k_3, k_4 \leq 2k_1, \quad (28.4)$$

и это дает множитель $2^{1/2}$ евклидовой норме. Очевидно, что этот результат очень слабый. Действительно, множитель $2^{-1/2}$ мог бы быть более реальным, чем $2^{1/2}$, потому что общая матрица ошибок, соответствующая обоим левому и правому умножениям, есть матрица вида

$$\left[\begin{array}{cc|cccc} p & q & \alpha_1 & \alpha_2 & \dots & \alpha_n \\ q & s & \beta_1 & \beta_2 & \dots & \beta_n \\ \hline \alpha_1 & \beta_1 & & & & \\ \alpha_2 & \beta_2 & & & & \\ \vdots & \vdots & & & & \\ \alpha_n & \beta_n & & & & \end{array} \right]. \quad (28.5)$$

В оценке § 27 мы считали отдельно евклидовы нормы вкладов строки и столбца. Если бы это не было сделано для элементов p, q, r, s , мы могли бы получить множитель $2^{-1/2}$ для совместного вклада.

Плоские вращения в арифметике с фиксированной запятой

29. Теперь обратим внимание на вычисления с фиксированной запятой. Так как каждый элемент ортогональной матрицы лежит в пределах от -1 до $+1$ и нормы преобразованных матриц остаются постоянными (с точностью до небольших изменений из-за ошибок округления), использование арифметики с фиксированной запятой не вызывает тяжелых проблем с нормировкой.

Рассмотрим сначала проблему § 20. Дан вектор $\bar{x}$ с компонентами с фиксированной запятой и мы хотим определить вращение в плоскости (i, j) так, чтобы $(R\bar{x})_j = 0$. Пусть

$$\sum \bar{x}_i^2 \leq 1. \quad (29.1)$$

При вычислении с плавающей запятой, используя уравнения (20.2), почти ортогональная матрица получается без каких-либо особых вычислительных хитростей, но так не всегда можно делать в арифметике с фиксированной запятой. Предположим, например, что мы имеем

$$\bar{x}_i = 0,000002, \quad \bar{x}_j = 0,000003, \quad (29.2)$$

и используем шестизначную десятичную арифметику. Тогда $fi(\bar{x}_i^2 + \bar{x}_j^2)^{1/2} = 0,000004$ (с точностью до шести десятичных знаков) и далее

$$\cos \theta = 0,500000, \quad \sin \theta = 0,750000. \quad (29.3)$$

Сразу же видно, что соответствующее приближенное плоское вращение далеко от ортогонального. Мы опишем два метода преодоления этого недостатка, первый из которых очень удобен для той вычислительной машины, в которой скалярное произведение может быть накоплено точно. В любом случае мы берем $\cos \theta = 1$ и $\sin \theta = 0$, если $\bar{x}_j = 0$. Обозначим

$$S^2 = \bar{x}_i^2 + \bar{x}_j^2, \quad (29.4)$$

и пусть S^2 вычислено точно; согласно (29.1) это не может дать переполнение. Затем определяем Σ^2 , где $\Sigma^2 = 2^{2k} S^2$ и k есть наименьшее целое, для которого

$$\frac{1}{4} < \Sigma^2 = 2^{2k} S^2 \leq 1. \quad (29.5)$$

Наконец, требуемые $\bar{s}$ и $\bar{c}$ вычисляются с использованием соотношений

$$\left. \begin{aligned} \bar{\Sigma} &= fi[(\Sigma^2)^{1/2}], \\ \bar{c} &= fi(2^k \bar{x}_i / \bar{\Sigma}), \quad \bar{s} = fi(2^k \bar{x}_j / \bar{\Sigma}). \end{aligned} \right\} \quad (29.6)$$

Величины $2^k \bar{x}_i$ и $2^k \bar{x}_j$ могут быть вычислены без ошибок округления, и мы имеем

$$\bar{\Sigma} \equiv \Sigma + \varepsilon_1, \quad |\varepsilon_1| \leq (1,00001) 2^{-t-1}, \quad (29.7)$$

$$\bar{c} \equiv \frac{2^k \bar{x}_i}{\bar{\Sigma}} + \varepsilon_2, \quad \bar{s} \equiv \frac{2^k \bar{x}_j}{\bar{\Sigma}} + \varepsilon_3, \quad (29.8)$$

$$|\varepsilon_i| \leq 2^{-t-1} \quad (i = 2, 3). \quad (29.9)$$

Объединяя эти результаты, получим

$$\bar{c} \equiv c + \varepsilon_4, \quad \bar{s} \equiv s + \varepsilon_5, \quad (29.10)$$

где

$$\varepsilon_4 = \frac{\varepsilon_1 c}{\bar{\Sigma} + \varepsilon_2}, \quad \varepsilon_5 = \frac{\varepsilon_1 s}{\bar{\Sigma} + \varepsilon_3}. \quad (29.11)$$

Нормировка S , $\bar{x}_i$ и $\bar{x}_j$ гарантирует, что $\bar{c}$ и $\bar{s}$ имеют малые абсолютные ошибки, хотя одно или другое может иметь большую относительную ошибку. Если мы запишем $\bar{R} = R + \delta R$, то δR — нулевая матрица, за исключением пересечения i и j строк и столбцов, где у нее будут такие элементы:

$$\begin{bmatrix} \varepsilon_4 & \varepsilon_5 \\ -\varepsilon_5 & \varepsilon_4 \end{bmatrix}. \quad (29.12)$$

Отсюда и из аддитивного свойства норм мы имеем

$$\|\delta R\|_2 \leq |\varepsilon_1| \bar{\Sigma} + (\varepsilon_2^2 + \varepsilon_3^2)^{1/2} < (1,71) 2^{-t} \quad (29.13)$$

для любого приемлемого t . Аналогично

$$\|\delta R\|_E < (1,71) 2^{1/2} 2^{-t} < (2,42) 2^{-t}. \quad (29.14)$$

Другое вычисление $\sin \theta$ и $\cos \theta$

30. Метод, который мы только что описали, на некоторых вычислительных машинах может быть не очень удобен. Следующая схема пригодна почти для любой машины.

Так как

$$\cos \theta = \frac{\bar{x}_i}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}}, \quad \sin \theta = \frac{\bar{x}_j}{(\bar{x}_i^2 + \bar{x}_j^2)^{1/2}}, \quad (30.1)$$

то если обозначим

$$\frac{\bar{x}_i}{\bar{x}_j} = m \quad \text{и} \quad \frac{\bar{x}_j}{\bar{x}_i} = n, \quad (30.2)$$

будем иметь

$$\cos \theta = \frac{m}{(m^2 + 1)^{1/2}} = \frac{1}{(n^2 + 1)^{1/2}}, \quad (30.3)$$

$$\sin \theta = \frac{1}{(m^2 + 1)^{1/2}} = \frac{n}{(n^2 + 1)^{1/2}}. \quad (30.4)$$

Мы можем использовать выражения, включающие m или n , смотря по тому, $|m|$ меньше или не меньше единицы. Для того чтобы держать все числа внутри допустимых пределов, (30.3) и (30.4) могут быть написаны

в виде

$$\cos \theta = \frac{\frac{1}{2} m}{\left(\frac{1}{4} m^2 + \frac{1}{4}\right)^{1/2}} = \frac{\frac{1}{2}}{\left(\frac{1}{4} n^2 + \frac{1}{4}\right)^{1/2}}, \quad (30.5)$$

$$\sin \theta = \frac{\frac{1}{2}}{\left(\frac{1}{4} m^2 + \frac{1}{4}\right)^{1/2}} = \frac{\frac{1}{2} n}{\left(\frac{1}{4} n^2 + \frac{1}{4}\right)^{1/2}}. \quad (30.6)$$

Применение этого метода опять дает значения $\bar{c}$ и $\bar{s}$, имеющие малые абсолютные ошибки.

Левое умножение на приближенное вращение с фиксированной запятой

31. Операция, рассмотренная в § 21, несколько проще при вычислениях с фиксированной запятой, и мы переходим немедленно к оценке $fi(\bar{R}A) - RA \equiv F$ без предварительного рассмотрения векторного умножения. Мы имеем

$$\begin{aligned} F \equiv fi(\bar{R}A) - RA &= fi(\bar{R}A) - \bar{R}A + \bar{R}A - RA = \\ &= fi(\bar{R}A) - \bar{R}A + \delta RA. \end{aligned} \quad (31.1)$$

Нам нужно сделать несколько предположений относительно величины элементов A , для того чтобы работа с фиксированной запятой была осуществима. Будем предполагать, что 2-нормы A и всех полученных матриц ограничены единицей и будем исследовать значение этого предположения только тогда, когда анализ будет завершен (ср. § 22).

При левом умножении изменяются лишь i -я и j -я строки, так что $fi(\bar{R}A) - \bar{R}A$ — нулевая матрица всюду, за исключением этих строк. Если мы предполагаем, что используем $fi_2(\)$ операции, то имеем

$$[fi_2(\bar{R}A)]_{ik} = fi_2(\bar{c}a_{ik} + \bar{s}a_{jk}) \equiv \bar{c}a_{ik} + \bar{s}a_{jk} + \varepsilon_{ik}, \quad (31.2)$$

$$[fi_2(\bar{R}A)]_{jk} = fi_2(-\bar{s}a_{ik} + \bar{c}a_{jk}) \equiv -\bar{s}a_{ik} + \bar{c}a_{jk} + \varepsilon_{jk}, \quad (31.3)$$

$$|\varepsilon_{ik}|, |\varepsilon_{jk}| \leq 2^{-t-1} \quad (i=1, \dots, n). \quad (31.4)$$

Для $fi(\)$ операций оценки вдвое больше. Соотношения (31.2) — (31.4) показывают, что мы можем написать

$$fi_2(\bar{R}A) - \bar{R}A \equiv G, \quad (31.5)$$

где G — нулевая, за исключением строк i и j , в которых она имеет элементы, ограниченные по модулю числом 2^{-t-1} . Следовательно,

$$\|G\|_2 \leq \|G\| \leq 2^{-t-1}(2n)^{1/2}, \quad (31.6)$$

$$\|G\|_E \leq 2^{-t-1}(2n)^{1/2}, \quad (31.7)$$

и (31.1) и (29.13) дают

$$\begin{aligned} \|F\|_2 = \|fi_2(\bar{R}A) - RA\|_2 &\leq \|G\|_2 + \|\delta R\|_2 \|A\|_2 \leq \\ &\leq 2^{-t-1}(2n)^{1/2} + (1,71) 2^{-t}. \end{aligned} \quad (31.8)$$

Заметим, что оценка вклада в ошибку, происходящую от действительного умножения, содержит множитель $n^{1/2}$, тогда как вклад, возникающий из-за отклонения $\bar{R}$ от R , не зависит от n . В вычислении с плавающей запятой оба вклада были почти равны и оба независимы от n . Заметим также, что даже если строки i и j целиком состоят из элементов, которые малы сравнительно с $\|A\|_2$, мы не получим более благоприятную оценку. Это существенно иной результат, чем тот, который мы получили для вычислений с плавающей запятой.

Если $\bar{x}_i$ и $\bar{x}_j$, которые определяют вращение, суть элементы некоторого столбца соответственно i -й и j -й строк $\bar{A}_{p-1}$, то обычно не вычисляют новых значений этих двух элементов по формулам (31.2) и (31.3). Вместо этого берутся значения $fi(2^{-k}\bar{\Sigma})$ (ср. § 29) и нуль. Нетрудно видеть, что это неопасно. Два элемента в RA равны соответственно $2^{-k}\Sigma$ и нулю, и, следовательно, ошибки в этих позициях будут $fi(2^{-k}\bar{\Sigma}) - 2^{-k}\Sigma$ и нуль. Мы имеем

$$fi(2^{-k}\bar{\Sigma}) - 2^{-k}\Sigma = 2^{-k}\bar{\Sigma} + \varepsilon - 2^{-k}\Sigma; \quad (|\varepsilon| \leq 2^{-t-1})$$

и, далее,

$$|fi(2^{-k}\bar{\Sigma}) - 2^{-k}\Sigma| \leq 2^{-k}|\bar{\Sigma} - \Sigma| + |\varepsilon| \leq 2^{-t-1}(1 + 2^{-k}) \leq 2^{-t}. \quad (31.9)$$

Ясно, что нам необходимо лишь изменить множитель в (31.8) с 1,71 на 2,21. Если теперь возьмем

$$\|F\|_2 \leq 2^{-t-1}(2n)^{1/2} + (2,21)2^{-t}, \quad (31.10)$$

то специальный случай покрывается.

Умножение на последовательность плоских вращений (фиксированная запятая)

32. Теперь рассмотрим вариант проблемы § 22 для фиксированной запятой. Аналогично уравнениям (22.1) и (22.2) имеем

$$\left. \begin{aligned} \bar{A}_1 &= fi_2(\bar{R}_1 A_0) \equiv R_1 A_0 + F_1, \\ \bar{A}_p &= fi_2(\bar{R}_p \bar{A}_{p-1}) \equiv R_p \bar{A}_{p-1} + F_p \quad (p = 2, \dots, s) \end{aligned} \right\} \quad (32.1)$$

и

$$\bar{A}_p = Q_1 A_0 + Q_2 F_1 + \dots + Q_p F_{p-1} + F_p, \quad (32.2)$$

где

$$Q_k = R_p R_{p-1} \dots R_k. \quad (32.3)$$

Следовательно,

$$\|\bar{A}_p - R_p R_{p-1} \dots R_1 A_0\|_2 \leq \|F_1\|_2 + \dots + \|F_p\|_2. \quad (32.4)$$

Если мы предполагаем $\|\bar{A}_i\|_2 \leq 1$ для всех преобразований, то анализ последнего параграфа показывает, что

$$\|F_i\|_2 \leq [(2n)^{1/2} + 4,42] 2^{-t-1} = x \quad (32.5)$$

и, следовательно,

$$\|\bar{A}_p - R_p R_{p-1} \dots R_1 A_0\|_2 \leq px, \quad (32.6)$$

откуда вытекает

$$\|\bar{A}_p\|_2 \leq \|A_0\|_2 + px. \quad (32.7)$$

Теперь мы можем утверждать, что если рассмотрено s преобразований, то при условии

$$\|A_0\|_2 < 1 - sx, \quad (32.8)$$

имеем

$$\|\bar{A}_p\|_2 < 1 \quad (p = 0, 1, \dots, s). \quad (32.9)$$

Для совокупности N преобразований так же, как было рассмотрено в § 22, наш окончательный результат таков:

$$\|\bar{A}_N - R_N R_{N-1} \dots R_1 A_0\|_2 \leq Nx \sim \frac{1}{4} n^{5/2} 2^{1/2} 2^{-t}, \quad (32.10)$$

если только A_0 нормирована так, что

$$\|A_0\|_2 \leq 1 - Nx. \quad (32.11)$$

33. Так как результат интересен для нас лишь тогда, когда Nx мало, то на практике требование первоначальной нормировки едва ли более обременительно, чем требование нормировки, которое было бы необходимо, если бы ошибки округления не имели места. Однако требование (32.11) непрактично, так как вычисление $\|A_0\|_2$ — такая же трудная проблема, как и вычисление собственных значений.

С другой стороны,

$$\|A_0\|_2 \leq \|A_0\|_E, \quad (33.1)$$

так что если

$$\|A_0\|_E < 1 - Nx, \quad (33.2)$$

то это же подавно верно для $\|A_0\|_2$, а $\|A_0\|_E$ легко вычисляется. Если первоначальная нормировка такова, что (33.2) удовлетворяется, то мы можем показать почти таким же путем, что

$$\|\bar{A}_N - R_N R_{N-1} \dots R_1 A_0\|_E \leq Ny, \quad (33.3)$$

где

$$y = [(2n)^{1/2} + 5,84] 2^{-t-1}. \quad (33.4)$$

Так как x и y почти равны, когда n велико, то если A_0 нормирована согласно (33.2), результат (33.3) будет более точным.

Вычисленное произведение приближенной последовательности плоских вращений

34. Мы можем применить эти результаты к вычислению произведения последовательности приближенных плоских вращений, как в § 25. Однако мы не можем брать $A_0 = I$ в (32.10), так как это нарушает (32.11). Если мы берем $A_0 = \frac{1}{2}I$, то (32.11) будет выполнено для любого значения n , для которого произведение имеет ценность. Если $\bar{P}$ — окончательно вычисленная матрица, то

$$\left\| \bar{P} - \frac{1}{2} R_N R_{N-1} \dots R_1 \right\|_2 \leq Nx, \quad (34.1)$$

$$\|2\bar{P} - R_N R_{N-1} \dots R_1\|_2 \leq 2Nx, \quad (34.2)$$

и ошибка опять порядка $n^{5/2}$.

Ошибки в подобных преобразованиях

35. Наш анализ немедленно дает оценки для ошибок, сделанных в последовательности N подобных преобразований. Действительно, если мы вычисляем

$$f_{i_2}(\bar{R}_N \bar{R}_{N-1} \dots \bar{R}_1 A_0 \bar{R}_1^T \dots \bar{R}_{N-1}^T \bar{R}_N^T),$$

перемножая матрицы в любом порядке, то

$$\| \bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T \|_2 \leq 2Nx \quad (35.1)$$

при условии, что

$$\| A_0 \|_2 \leq 1 - 2Nx. \quad (35.2)$$

Однако пары чисел, которые определяют $\bar{R}_p$, часто будут элементами $\bar{A}_{p-1}$, и в этом случае вычисления будут выполняться в следующем порядке:

$$\left. \begin{aligned} \bar{X}_1 &= f_{i_2}(\bar{R}_1 A_0), & \bar{A}_1 &= f_{i_2}(\bar{X}_1 \bar{R}_1^T), \\ \bar{X}_p &= f_{i_2}(\bar{R}_p \bar{A}_{p-1}), & \bar{A}_p &= f_{i_2}(\bar{X}_p \bar{R}_p^T) \quad (p = 2, \dots, N). \end{aligned} \right\} \quad (35.3)$$

В этом случае выгодно рассмотреть совместно левое и правое умножения. Мы определим H_p соотношением

$$H_p = \bar{A}_p - \bar{R}_p \bar{A}_{p-1} \bar{R}_p^T = f_{i_2}(\bar{R}_p \bar{A}_{p-1} \bar{R}_p^T) - \bar{R}_p \bar{A}_{p-1} \bar{R}_p^T, \quad (35.4)$$

так что

$$\begin{aligned} \bar{A}_p - R_p \bar{A}_{p-1} R_p^T &= H_p + \bar{R}_p \bar{A}_{p-1} \bar{R}_p^T - R_p \bar{A}_{p-1} R_p^T = \\ &= H_p + \delta R_p \bar{A}_{p-1} R_p^T + R_p \bar{A}_{p-1} \delta R_p^T + \delta R_p \bar{A}_{p-1} \delta R_p^T. \end{aligned} \quad (35.5)$$

Следовательно,

$$\begin{aligned} \| \bar{A}_p - R_p \bar{A}_{p-1} R_p^T \|_2 &\leq \| H_p \|_2 + (2 \| \delta R_p \|_2 + \| \delta R_p \|_2^2) \| \bar{A}_{p-1} \|_2 \leq \\ &\leq \| H_p \|_2 + (3,43) 2^{-t} \| \bar{A}_{p-1} \|_2, \end{aligned} \quad (35.6)$$

где мы заменили $[2(1,71) 2^{-t} + (1,71)^2 \cdot 2^{-2t}]$ на $(3,43) 2^{-t}$. При условии $\| \bar{A}_{p-1} \|_2 < 1$ для всех соответствующих p мы имеем теперь

$$\| \bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T \|_2 \leq \sum_1^N \| H_p \|_2 + N(3,43) 2^{-t}. \quad (35.7)$$

Здесь H_p — матрица ошибок, сделанных при вычислении $\bar{A}_p$ из $\bar{R}_p \bar{A}_{p-1} \bar{R}_p^T$. Если вращение в плоскости (i, j) , то преобразование влияет лишь на элементы i -х и j -х строк и столбцов и, следовательно, в остальных местах H_p нулевая. Помимо четырех элементов, стоящих на пересечении, при левом умножении изменяются лишь строки i и j , а при правом — столбцы i и j . Поэтому имеем

$$(H_p)_{ik} = f_{i_2}(\bar{a}_{ik} \bar{c} + \bar{a}_{jk} \bar{s}) - \bar{a}_{ik} \bar{c} - \bar{a}_{jk} \bar{s} = \varepsilon_{ik} \quad (|\varepsilon_{ik}| \leq 2^{-t-1}) \quad (35.8)$$

для всех $k \neq i, j$. Аналогично

$$|(H_p)_{jk}|, |(H_p)_{ki}|, |(H_p)_{kj}| \leq 2^{-t-1}. \quad (35.9)$$

Элементы четырех точек пересечения изменяются как при левом, так и при правом умножении, и мы их сейчас рассмотрим подробно. Предположим, что сначала выполняется левое умножение. Ошибки, сделанные при левом

умножении в этих четырех позициях, будут такими же по величине, как и ошибки, сделанные в остальных элементах строк i и j . Поэтому мы можем обозначить полученные значения через

$$\begin{bmatrix} A + \varepsilon_{ii} & B + \varepsilon_{ij} \\ C + \varepsilon_{ji} & D + \varepsilon_{jj} \end{bmatrix} = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}, \quad (35.10)$$

где A, B, C, D соответствуют точному левому умножению на $\bar{R}_p$. Далее, значения, полученные после правого умножения, могут быть выражены в виде

$$\begin{bmatrix} \bar{A}\bar{c} + \bar{B}\bar{s} + \eta_{ii} & -\bar{A}\bar{s} + \bar{B}\bar{c} + \eta_{ij} \\ \bar{C}\bar{c} + \bar{D}\bar{s} + \eta_{ji} & -\bar{C}\bar{s} + \bar{D}\bar{c} + \eta_{jj} \end{bmatrix}, \quad (35.11)$$

где каждое η_{st} удовлетворяет неравенству

$$|\eta_{st}| \leq 2^{-t-1}. \quad (35.12)$$

Если мы обозначим ошибки в этих позициях при смешанной операции (относительно $\bar{R}_p \bar{A}_p^{-1} \bar{R}_p^T$) через

$$2^{-t-1} \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \quad (35.13)$$

то

$$2^{-t-1} \alpha = \varepsilon_{ii} \bar{c} + \varepsilon_{ij} \bar{s} + \eta_{ii},$$

$$2^{-t-1} |\alpha| \leq |\varepsilon_{ii}| |\bar{c}| + |\varepsilon_{ij}| |\bar{s}| + |\eta_{ii}| \leq (\varepsilon_{ii}^2 + \varepsilon_{ij}^2)^{1/2} (\bar{c}^2 + \bar{s}^2)^{1/2} + |\eta_{ii}| < (2,42) 2^{-t-1}. \quad (35.14)$$

Эта же оценка, очевидно, пригодна для β, γ, δ и, следовательно, например, при $n = 6, i = 2, j = 4$:

$$|H_p| \leq 2^{-t-1} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 2,42 & 1 & 2,42 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 2,42 & 1 & 2,42 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \quad (35.15)$$

Здесь симметричная матрица, образованная из единичных элементов, имеет характеристическое уравнение

$$\lambda^n - 2(n-2)\lambda^{n-2} = 0, \quad (35.16)$$

и, следовательно, ее 2-норма есть $(2n-4)^{1/2}$; 2-норма оставшейся матрицы из четырех элементов не больше чем $(4,84) 2^{-t-1}$.

Объединяя эти результаты, получим

$$\|H_p\|_2 < [(2n-4)^{1/2} + 4,84] 2^{-t-1}, \quad (35.17)$$

и поэтому из (35.7)

$$\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_2 \leq 2^{-t-1} [(2n-4)^{1/2} + 11,70] N. \quad (35.18)$$

Соответствующее требование первоначальной нормировки таково:

$$\|A_0\|_2 < 1 - 2^{-t-1} [(2n-4)^{1/2} + 11,70] N. \quad (35.19)$$

Общее замечание об оценках ошибки

36. Оценки для фиксированной запятой, которые мы только что получили, не имеют экспоненциального множителя того типа, который мы имели в оценках для плавающей запятой. Однако это различие обманчиво, так как мы должны помнить, что наша оценка имеет место только в предположении, что (35.19) выполнено. Если n очень велико, это условие не может быть удовлетворено, так как оно означает, что $\|A_0\|_2$ должна быть отрицательной. В пределах своего использования все результаты в основном такого вида:

$$\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_E \leq K_1 n^{3/2} 2^{-t} \|A_0\|_E - \text{плавающая запятая,} \quad (36.1)$$

$$\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_2 \leq K_2 n^{5/2} 2^{-t} \|A_0\|_2 - \text{фиксированная} \\ \text{запятая,} \quad (36.2)$$

$$\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_E \leq K_3 n^{5/2} 2^{-t} \|A_0\|_E - \text{фиксированная} \\ \text{запятая.} \quad (36.3)$$

Нет никакого сомнения, что результаты для плавающей запятой существенно точнее для больших n .

Возможно, имеет смысл напомнить, что маловероятно, чтобы результат для фиксированной запятой и общей матрицы A_0 мог бы быть улучшен в множителе $n^{5/2} 2^{-t}$. Рассмотрим, например, следующий гипотетический пример, в котором матрицы $\bar{A}_p$ ($p = 1, \dots, N$) вычисляем следующим образом.

На p -м шаге матрицы R_p и $R_p \bar{A}_{p-1} R_p^T$ вычисляются точно с использованием бесконечного числа знаков, и после этого точного вычисления элементы $R_p \bar{A}_{p-1} R_p^T$ округляются до стандартных чисел с фиксированной запятой. Полученная матрица определяется как $\bar{A}_p$. Очевидно, что

$$\bar{A}_p = R_p \bar{A}_{p-1} R_p^T + H_p, \quad (36.4)$$

где $|H_p|$ ограничена для $n = 6$, $i = 2$, $j = 4$ матрицей

$$2^{-t-1} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \quad (36.5)$$

(Мы предполагаем здесь общий случай, так что нет нулевых элементов, полученных в $\bar{A}_p$ при преобразовании.) Теперь отсюда следует, как и в § 35, что даже для этого примера оценка, полученная для $\|\bar{A}_N - R_N \dots R_1 A_0 R_1^T \dots R_N^T\|_2$, имеет множитель $n^{5/2} 2^{-t}$.

Использование время от времени операций $fi_2(\)$ вместо $fi(\)$ оказывает малое влияние на окончательную оценку, требуется лишь дополнительный множитель 2, если всюду используется $fi(\)$. В анализе с плавающей запятой мы всюду использовали $fl(\)$, но использование $fl_2(\)$ не должно было внести существенного различия, так как во всем вычислении нигде не встречались скалярные произведения большого порядка.

Матрицы отражения с плавающей запятой

37. Мы обратимся теперь к использованию матриц отражения (гл. 1, §§ 45, 46). Нецелесообразно давать анализ каждого из fi (), fi_2 (), fl () и fl_2 () способов. Мы выбрали fl_2 () способ, потому что он эффективен на практике, а анализ остальных способов мы дали в других работах (см., например, Уилкинсон (1961b) и (1962b)). В самом общем случае положение, возникающее в существующих алгоритмах, состоит в следующем.

Пусть нам дан вектор x порядка n с компонентами с плавающей запятой, и мы хотим определить матрицу отражения P так, чтобы левое умножение на P оставляло первые r компонент x без изменения и исключало элементы $r + 2, r + 3, \dots, n$. Будем предполагать, что x — вещественный, и дадим соответствующий математический анализ, не учитывая вначале ошибки округления.

Если мы напишем

$$w^T = (0, \dots, 0, w_{r+1}, \dots, w_n) = \quad (37.1)$$

$$= (0 \mid v^T), \quad (37.2)$$

где v — вектор с $(n - r)$ компонентами, то

$$P = I - 2ww^T = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2vv^T \end{array} \right]. \quad (37.3)$$

Здесь $(I - 2vv^T)$ — матрица отражения порядка $(n - r)$, так как

$$v^T v = w^T w = 1. \quad (37.4)$$

Представив x^T для согласования с P в виде

$$x^T = (y^T \mid z^T), \quad (37.5)$$

имеем

$$Px = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2vv^T \end{array} \right] \left[\begin{array}{c} y \\ z \end{array} \right] = \left[\begin{array}{c} y \\ \frac{y}{(I - 2vv^T)z} \end{array} \right]. \quad (37.6)$$

Следовательно, мы должны выбрать v только так, чтобы $(I - 2vv^T)z$ имел нули везде, кроме своей первой компоненты.

38. Очевидно, что наша исходная задача, которая имела параметры r и n , в основном эквивалентна более простой задаче с параметрами 0 и $n - r$. Поэтому мы не потеряем общности, если временно рассмотрим следующую проблему. Дан вектор x порядка n , нужно построить матрицу отражения такую, что

$$Px = ke_1. \quad (38.1)$$

Так как 2-норма инвариантна, то если

$$S^2 = x_1^2 + \dots + x_n^2, \quad (38.2)$$

будем иметь

$$k = \pm S. \quad (38.3)$$

Поэтому уравнение (38.1) дает

$$x_1 - 2w_1(w^T x) = \pm S, \quad (38.4)$$

$$x_i - 2w_i(w^T x) = 0 \quad (i = 2, \dots, n) \quad (38.5)$$

и, следовательно,

$$2Kw_1 = x_1 \mp S, \quad 2Kw_i = x_i \quad (i = 2, \dots, n), \quad (38.6)$$

где

$$K = w^T x. \quad (38.7)$$

Возводя уравнения (38.6) в квадрат, складывая и деля на два, получаем

$$2K^2 = S^2 \mp x_1 S. \quad (38.8)$$

Если мы напомним

$$u^T = (x_1 \mp S, x_2, x_3, \dots, x_n), \quad (38.9)$$

то

$$w = \frac{u}{2K}, \quad (38.10)$$

и

$$P = I - \frac{uu^T}{2K^2}. \quad (38.11)$$

Во всем процессе содержится лишь одно извлечение квадратного корня при получении S из S^2 .

Анализ ошибок вычисления матрицы отражения

39. Предположим теперь, что x имеет компоненты с плавающей запятой и используем $fl_2(\quad)$ вычисление, чтобы получить матрицу $\bar{P}$. Мы хотим найти оценку для $\|\bar{P} - P\|$, где P — точная матрица, соответствующая данному x . Так как для целей данного вычисления можно рассматривать x как точный вектор, то мы опустим черточки над его компонентами.

В уравнениях § 38 есть свобода в выборе знака. Мы найдем, что для численной устойчивости правильный выбор таков, что

$$2K^2 = S^2 \mp x_1 S = S^2 + |x_1| S, \quad (39.1)$$

так что никакое уничтожение знаков при вычислении $2K^2$ не имеет места. Итак, если x_1 положительно, берем

$$2K^2 = S^2 + x_1 S, \quad (39.2)$$

$$u^T = (x_1 + S, x_2, \dots, x_n). \quad (39.3)$$

В предположении, что x_1 должно быть положительно, не потеряно никакой общности; при условии, что мы берем устойчивый выбор знака, оценки ошибки не зависят от знака x_1 .

Этапы вычисления состоят в следующем.

(i) Вычисляем величину a , определенную равенством

$$a = fl_2(x_1^2 + x_2^2 + \dots + x_n^2), \quad (39.4)$$

причем сохраняем в a $2t$ -разрядную мантиссу, так как она требуется на этапах (ii) и (iii). Итак,

$$a \equiv x_1^2(1 + \varepsilon_1) + \dots + x_n^2(1 + \varepsilon_n), \quad (39.5)$$

и заведомо из (9.4) и (9.5) имеем

$$|\varepsilon_i| < \frac{3}{2} n 2^{-2t_2}. \quad (39.6)$$

Поэтому

$$a \equiv (x_1^2 + x_2^2 + \dots + x_n^2) (1 + \varepsilon) = S^2 (1 + \varepsilon); \quad \left(|\varepsilon| < \frac{3}{2} n 2^{-2t_2} \right). \quad (39.7)$$

(ii) По a вычисляем $\bar{S}$, где

$$\bar{S} = fl_2(a^{1/2}) \equiv a^{1/2} (1 + \eta) \quad (|\eta| < (1,00001) 2^{-t}) \quad (39.8)$$

с учетом (10.2). Следовательно,

$$\bar{S} \equiv S (1 + \varepsilon)^{1/2} (1 + \eta) \equiv S (1 + \xi), \quad (39.9)$$

где по (39.7) и (39.8)

$$\xi < (1,00002) 2^{-t}. \quad (39.10)$$

(iii) Вычисляем $2\bar{K}^2$, где

$$\begin{aligned} 2\bar{K}^2 &= fl_2(x_1^2 + x_2^2 + \dots + x_n^2 + x_1\bar{S}) \equiv \\ &\equiv [x_1^2(1 + \eta_1) + x_2^2(1 + \eta_2) + \dots + x_n^2(1 + \eta_n) + x_1\bar{S}(1 + \eta_{n+1})] (1 + \varepsilon), \end{aligned} \quad (39.11)$$

причем заведомо из (9.4) и (9.5)

$$|\eta_i| < \frac{3}{2} (n+1) 2^{-2t_2}, \quad |\varepsilon| \leq 2^{-t}. \quad (39.12)$$

В этом вычислении $2\bar{K}^2$ большая часть накопления скалярного произведения была сделана в (i).

Объединяя эти результаты и помня, что x_1 положительно, имеем

$$\begin{aligned} 2\bar{K}^2 &\equiv (x_1^2 + x_2^2 + \dots + x_n^2 + x_1\bar{S}) (1 + \xi) (1 + \varepsilon) \equiv \\ &\equiv [S^2 + x_1 S (1 + \xi)] (1 + \xi) (1 + \varepsilon) \left(|\xi| < \frac{3}{2} (n+1) 2^{-2t_2} \right). \end{aligned} \quad (39.13)$$

Теперь, если мы запишем

$$S^2 + x_1 S (1 + \xi) \equiv (S^2 + x_1 S) (1 + k), \quad (39.14)$$

то будем иметь

$$k = \left(\frac{x_1 S}{S^2 + x_1 S} \right) \xi. \quad (39.15)$$

Так как $x_1 S$ и S^2 положительны и $x_1 \leq S$, то это дает

$$|k| \leq \frac{1}{2} |\xi|. \quad (39.16)$$

Опять важен выбор знака; если x_1 отрицательно, то мы должны взять $S^2 - x_1 S$, чтобы быть уверенными в малой относительной ошибке. Поэтому окончательно

$$2\bar{K}^2 \equiv 2K^2 (1 + \theta) \quad (|\theta| < (1,501) 2^{-t}). \quad (39.17)$$

(iv) Наконец, вычисляем $\bar{u}$. Лишь один первый элемент нуждается в вычислении, и мы имеем

$$\bar{u}_1 = fl(x_1 + \bar{S}) \equiv (x_1 + \bar{S}) (1 + \varphi) \equiv \quad (39.18)$$

$$\equiv [x_1 + S (1 + \xi)] (1 + \varphi) \quad (|\varphi| \leq 2^{-t}). \quad (39.19)$$

Наш выбор знака также гарантирует здесь малую относительную ошибку поэтому

$$\bar{u}_1 \equiv (x_1 + S)(1 + \psi) \equiv \quad (39.20)$$

$$\equiv u_1(1 + \psi) \quad (|\psi| < (1,501) 2^{-t}). \quad (39.21)$$

Мы можем написать

$$\bar{u}^T \equiv (\bar{u}_1, u_2, u_3, \dots, u_n) \equiv u^T + \delta u^T, \quad (39.22)$$

где

$$\|\delta u\|_2 = |u_1| \cdot |\psi| < (1,501) 2^{-t} \|u\|_2, \quad (39.23)$$

и, следовательно,

$$\|\bar{u}\|_2 = \|u + \delta u\|_2 < [1 + (1,501) 2^{-t}] \|u\|_2. \quad (39.24)$$

40. Матрицу P выгодно оставить в факторизованной форме

$$\bar{P} = I - \frac{\bar{u}\bar{u}^T}{2\bar{K}^2}, \quad (40.1)$$

так как часто такое представление будет самым удобным для работы. Никакие дальнейшие ошибки округления не появляются до тех пор, пока мы не попытаемся использовать $\bar{P}$ как множитель, а это требует отдельного анализа. Из (40.1)

$$\bar{P} - P = uu^T/2K^2 - \bar{u}\bar{u}^T/2\bar{K}^2 =$$

$$= (uu^T - \bar{u}\bar{u}^T)/2K^2 + \bar{u}\bar{u}^T \left(\frac{1}{2} K^2 - \frac{1}{2} \bar{K}^2 \right). \quad (40.2)$$

Так как $\|uu^T/2K^2\|_2 = 2$, то простое вычисление, использующее (39.17), (39.23) и (39.24), показывает, что

$$\|\bar{P} - P\|_2 < (9,01) 2^{-t}. \quad (40.3)$$

Отметим, что оценка не зависит от n и, так как P эквивалентна $(n-1)$ плоским вращениям, весьма хороша.

Для некоторых приложений удобно преобразовать P и вычислить вектор $\bar{v}$, определенный равенством

$$\bar{v}^T = fl \left(\frac{\bar{u}^T}{2\bar{K}^2} \right), \quad (40.4)$$

так что

$$\bar{P} = I - \bar{u}\bar{v}^T. \quad (40.5)$$

Из (40.4)

$$\bar{v}_i^T \equiv \frac{\bar{u}_i^T}{2\bar{K}^2} (1 + \epsilon_i) \quad (|\epsilon_i| \leq 2^{-t}) \quad (40.6)$$

и, следовательно,

$$\bar{v} \equiv \frac{\bar{u}}{2\bar{K}^2} + \delta v, \quad (40.7)$$

где

$$\|\delta v\|_2 \leq \frac{2^{-t} \|\bar{u}\|_2}{2\bar{K}^2}. \quad (40.8)$$

Для $\bar{P}$, вычисленной таким образом, мы имеем

$$\|\bar{P} - P\|_2 < (11,02) 2^{-t}. \quad (40.9)$$

Так как $\|ab^T\|_2 = \|ab^T\|_E$ для любых векторов a и b , то мы можем заменить 2-норму евклидовой в приведенных выше оценках.

Возвращаясь к случаю, когда изменяются лишь последние $(n - r)$ компонент, мы получим матрицу вида

$$\left[\begin{array}{c|c} I & 0 \\ \hline 0 & \bar{P} \end{array} \right],$$

где $\bar{P}$ порядка $(n - r)$. Напомним, однако, что наша оценка ошибки не зависит от n .

Численный пример

41. Чрезвычайная важность правильного выбора знака может быть проиллюстрирована следующим простым численным примером. Предположим, что задан вектор x :

$$x = [10^0(0,5123); 10^{-3}(0,6147); 10^{-3}(0,5135)].$$

Используя операцию $fl_2(\quad)$, соответствующую 4-разрядной десятичной арифметике, имеем

$$\begin{aligned} a &= fl_2(x_1^2 + x_2^2 + x_3^2) \equiv 10^0(0,26245193), \\ \bar{S} &= fl_2(a^{1/2}) \equiv 10^0(0,5123), \end{aligned}$$

и при устойчивом выборе знака

$$\begin{aligned} 2\bar{K}^2 &= fl_2(x_1^2 + x_2^2 + x_3^2 + x_1\bar{S}) = 10^0(0,5249), \\ \bar{u}_1 &= fl_2(x_1 + \bar{S}) = 10^1(0,1025), \\ \bar{u}^T &= [10^1(0,1025); 10^{-3}(0,6147); 10^{-3}(0,5135)]. \end{aligned}$$

«Правильные» значения с этим выбором знака таковы:

$$\begin{aligned} S &= 10^0(0,5123006261...), \quad 2K^2 = 10^0(0,5249035...), \\ u^T &= [10^1(0,10246006...); 10^{-3}(0,6147); 10^{-3}(0,5135)]. \end{aligned}$$

Можно проверить, что значения с чертой удовлетворяют неравенствам § 39, где мы заменяем 2^{-t} на $\frac{1}{2} \cdot 10^{-3}$ соответственно точности четырех-значных десятичных вычислений.

Используя неустойчивый выбор знака, имеем

$$\begin{aligned} 2\bar{K}^2 &= fl_2(x_1^2 + x_2^2 + x_3^2 - x_1\bar{S}) \equiv 10^{-6}(0,6400), \\ \bar{u}_1 &= fl_2(x_1 - \bar{S}) \equiv 10^0(0,0000), \\ \bar{u}^T &= [10^0(0,0000); 10^{-3}(0,6147); 10^{-3}(0,5135)]. \end{aligned}$$

«Правильные» значения с таким выбором знака будут

$$2K^2 = 10^{-6}(0,3207...), \quad u^T = [10^{-6}(-0,6264); 10^{-3}(0,6147); 10^{-3}(0,5135)].$$

Легко может быть проверено, что $\bar{P} = I - \bar{u}\bar{u}^T/2\bar{K}^2$ в этом случае даже приближенно неортогональна и поэтому не близка к P . Вычисленное значение $2\bar{K}^2$ имеет очень большую относительную ошибку.

Левое умножение на приближенную матрицу отражения

42. По аналогии с исследованием плоских вращений нам теперь нужна оценка для

$$fl_2 \left\{ \left[\begin{array}{c|c} I & 0 \\ \hline 0 & P \end{array} \right] A \right\} - \left[\begin{array}{c|c} I & 0 \\ \hline 0 & P \end{array} \right] A. \quad (42.1)$$

Очевидно, что левое умножение оставляет первые r строк A неизменными, а последние $(n - r)$ строк умножаются на P (или на $\bar{P}$ с ошибками округления). Поэтому матрица ошибок (42.1) нулевая в своих первых r строках, и, не теряя общности, можно рассматривать левое умножение матрицы A из m строк и n столбцов на матрицу P , имеющую для своего соответствующего w вектор порядка m с ненулевыми компонентами.

Мы берем $\bar{P}$ в факторизованной форме $(I - \bar{u}\bar{u}^T/2\bar{K}^2)$, так что неравенства § 39 и (40.3) удовлетворяются.

Имеем

$$fl_2(\bar{P}A) - PA = fl_2(\bar{P}A) - \bar{P}A + \bar{P}A - PA, \quad (42.2)$$

$$\begin{aligned} \|fl_2(\bar{P}A) - PA\|_E &\leq \|fl_2(\bar{P}A) - \bar{P}A\|_E + \|(\bar{P} - P)A\|_E \leq \\ &\leq \|fl_2(\bar{P}A) - \bar{P}A\|_E + (9,01)2^{-t}\|A\|_E. \end{aligned} \quad (42.3)$$

Итак, нам нужна оценка для матрицы $fl_2(\bar{P}A) - \bar{P}A$, которая является матрицей ошибок, действительно возникающих при левом умножении. Заметим, что оценка ошибки, которую мы дали в (9.14), неприменима к вычислению $fl_2(\bar{P}A)$, так как для экономии объема вычислений $\bar{P}$ используется в факторизованной форме. Для того чтобы подчеркнуть тот факт, что мы имеем дело лишь с ошибками, сделанными при левом умножении, введем y и M , определенные равенствами

$$y = \bar{u}, \quad M = 2\bar{K}^2, \quad (42.4)$$

так что

$$\bar{P} = I - \frac{yy^T}{M} \quad (42.5)$$

и

$$\begin{aligned} \|y\|_2^2/M &= \|\bar{u}\|_2^2/2\bar{K}^2 < \|u\|_2^2[1 + (1,501)2^{-t}]^2/2K^2[1 - (1,501)2^{-t}] < \\ &< 2[1 + (4,51)2^{-t}]. \end{aligned} \quad (42.6)$$

43. Мы можем выделить этапы в вычислении точного $\bar{P}A$ и приближенного $fl_2(PA)$ следующим образом:

Точное	Приближенное	
$p^T = y^T A/M;$	$\bar{p}^T = fl_2(y^T A/M);$	(43.1)
$B = \bar{P}A = A - yp^T;$	$\bar{B} = fl_2(\bar{P}A) = fl_2(A - y\bar{p}^T).$	(43.2)

Поэтому i -я компонента $\bar{p}_i$ дается равенством

$$\bar{p}_i \equiv [y_1 a_{1i} (1 + \varepsilon_1) + y_2 a_{2i} (1 + \varepsilon_2) + \dots + y_m a_{mi} (1 + \varepsilon_m)] (1 + \varepsilon) / M, \quad (43.3)$$

где из (9.2) и (9.5) имеем заведомо

$$|\varepsilon_i| < \frac{3}{2} (m+1) 2^{-2t_2}, \quad |\varepsilon| \leq 2^{-t}. \quad (43.4)$$

Следовательно,

$$\bar{p}_i = p_i (1 + \varepsilon) + q_i, \quad (43.5)$$

где

$$|q_i| \leq \frac{3}{2} (m+1) 2^{-2t_2} (1 + 2^{-t}) [|y_1| |a_{1i}| + |y_2| |a_{2i}| + \dots + |y_m| |a_{mi}|] / M. \quad (43.6)$$

Отсюда

$$|\bar{p}_i - p_i| \leq 2^{-t} |p_i| + \frac{3}{2} (m+1) 2^{-2t_2} (1 + 2^{-t}) [|A|^T |y|]_i / M. \quad (43.7)$$

Итак,

$$\|\bar{p} - p\|_2 \leq 2^{-t} \|p\|_2 + \frac{3}{2} (m+1) 2^{-2t_2} (1 + 2^{-t}) \|y\|_2 \|A\|_E / M, \quad (43.8)$$

и, принимая во внимание предположение (5.2), что $n \cdot 2^{-t} < 0,4$, мы можем написать

$$\|\bar{p} - p\|_2 \leq 2^{-t} \|p\|_2 + (0,16) 2^{-t} \|y\|_2 \|A\|_E / M. \quad (43.9)$$

Для компоненты (i, j) матрицы $\bar{B}$ справедливы соотношения

$$\bar{b}_{ij} = fl_2(A - yp^T)_{ij} \equiv [a_{ij}(1 + \eta_1) - y_i \bar{p}_j (1 + \eta_2)] (1 + \xi), \quad (43.10)$$

$$|\eta_1|, |\eta_2| \leq \frac{3}{2} 2^{-2t}, \quad |\xi| \leq 2^{-t}. \quad (43.11)$$

Если мы определим δp равенством

$$\bar{p} = p + \delta p, \quad (43.12)$$

то

$$\begin{aligned} \bar{b}_{ij} &\equiv [a_{ij}(1 + \eta_1) - y_i(p_j + \delta p_j)(1 + \eta_2)](1 + \xi) = b_{ij}(1 + \xi) + \\ &+ [(a_{ij}\eta_1 - y_i p_j \eta_2 - y_i \delta p_j (1 + \eta_2)](1 + \xi). \end{aligned} \quad (43.13)$$

Следовательно,

$$\begin{aligned} |\bar{b}_{ij} - b_{ij}| &\leq 2^{-t} |b_{ij}| + \\ &+ \left[\frac{3}{2} 2^{-2t} |a_{ij}| + \frac{3}{2} 2^{-2t} |y_i| |p_j| + |y_i| |\delta p_j| \left(1 + \frac{3}{2} 2^{-2t} \right) \right] (1 + 2^{-t}), \end{aligned} \quad (43.14)$$

поэтому

$$\begin{aligned} |\bar{B} - B| &\leq 2^{-t} |B| + \\ &+ \left[\frac{3}{2} 2^{-2t} |A| + \frac{3}{2} 2^{-2t} \|y\| \|p\|^T + \|y\| \|\delta p\|^T \left(1 + \frac{3}{2} 2^{-2t} \right) \right] (1 + 2^{-t}), \end{aligned} \quad (43.15)$$

откуда выводим

$$\|\bar{B} - B\|_E \leq 2^{-t} \|B\|_E + \left[\frac{3}{2} 2^{-2t} \|A\|_E + \frac{3}{2} 2^{-2t} \|y\|_2 \|P\|_2 + \right. \\ \left. + \|y\|_2 \|\delta P\|_2 \left(1 + \frac{3}{2} 2^{-2t} \right) \right] (1 + 2^{-t}). \quad (43.16)$$

44. Теперь нам нужны оценки для каждого члена правой части через $\|A\|_E$. Из (43.1) и (40.3) вытекает

$$\|B\|_E = \|\bar{P}A\|_E \leq \|PA\|_E + \|(\bar{P} - P)A\|_E \leq \|A\|_E + (9,01) 2^{-t} \|A\|_E, \quad (44.1)$$

из (43.1) и (42.6) —

$$\|y\|_2 \|P\|_2 \leq \|y\|_2^2 \|A\|_E / M < 2 \|A\|_E [1 + (4,51) 2^{-t}], \quad (44.2)$$

из (43.9) и (42.6) —

$$\|y\|_2 \|\delta P\|_2 \leq \|y\|_2 [2^{-t} \|P\|_2 + (0,16) 2^{-t} \|y\|_2 \|A\|_E / M] \leq \\ \leq 2^{-t} \{2 \|A\|_E [1 + (4,51) 2^{-t}] + (0,32) \|A\|_E [1 + (4,51) 2^{-t}]\} \leq \\ \leq (2,33) 2^{-t} \|A\|_E. \quad (44.3)$$

Объединяя эти результаты, имеем

$$\|\bar{B} - B\|_E \leq (3,35) 2^{-t} \|A\|_E, \quad (44.4)$$

и вспоминая, что $\bar{B} - B = fl_2(\bar{P}A) - \bar{P}A$, из соотношения (42.3) получаем

$$\|fl_2(\bar{P}A) - \bar{P}A\|_E \leq (3,35) 2^{-t} \|A\|_E + (9,01) 2^{-t} \|A\|_E = (12,36) 2^{-t} \|A\|_E. \quad (44.5)$$

Возвращаясь к общему случаю, непосредственно получаем

$$\left\| fl_2 \left\{ \left[\begin{array}{c|c} I & 0 \\ \hline 0 & P \end{array} \right] A \right\} - \left[\begin{array}{c|c} I & 0 \\ \hline 0 & P \end{array} \right] A \right\|_E \leq \\ \leq (12,36) 2^{-t} \|A_{n-r}\|_E \leq (12,36) 2^{-t} \|A\|_E, \quad (44.6)$$

где A_{n-r} есть матрица, состоящая из последних $(n-r)$ строк A . На практике, когда один из столбцов A является вектором x , по которому определяется P , мы не вычисляем в действительности элементы Px , а автоматически приписываем соответствующее число нулевых элементов и даем оставшемуся элементу значение $\pm \bar{S}$. Соответствующие значения в PA будут нуль и $\pm S$, так что единственный вклад в матрицу ошибок согласно (39.9) есть элемент $\pm S\xi$. Поэтому специальный прием для такого столбца покрывается нашим общим анализом.

Умножение на последовательность приближенных матриц отражения

45. В большом числе алгоритмов мы будем иметь дело с последовательностью из $(n-1)$ матрицы отражения $P_0, P_1, \dots, P_{n-2}$, где вектор w_r , соответствующий P_r , имеет нулевые элементы в первых r позициях. Мы сначала рассмотрим левые умножения и обозначим вычисленные P_r через

$\bar{P}_r$, а вычисленные преобразования через $\bar{A}_r$. Тогда имеем

$$\bar{A}_1 - P_0 A_0 = F_0, \quad \|F_0\|_E \leq (12,36) 2^{-t} \|A_0\|_E, \quad (45.1)$$

$$\bar{A}_{r+1} - P_r \bar{A}_r = F_r, \quad \|F_r\|_E \leq (12,36) 2^{-t} \|\bar{A}_r\|_E, \quad (45.2)$$

и, следовательно, если обозначим $(12,36) 2^{-t}$ через x , то

$$\begin{aligned} \|\bar{A}_{n-1} - P_{n-2} P_{n-3} \dots P_0 A_0\|_E &\leq x[1 + (1+x) + (1+x)^2 + \dots + \\ &+ (1+x)^{n-2}] \|A_0\|_E \leq (n-1)(1+x)^{n-2} x \|A_0\|_E. \end{aligned} \quad (45.3)$$

Результаты будут интересны для нас лишь тогда, когда nx существенно меньше единицы, так что, хотя множитель $(1+x)^{n-2}$ и растет экспоненциально, он оказывает слабое влияние на возможный предел использования этой формулы. По существу наша оценка имеет вид $12,36n2^{-t} \|A_0\|_E$, и она замечательна, особенно если вспомнить, что мы не принимали во внимание статистическое распределение ошибок.

Заметим, что наше исследование слабо в некотором отношении, так как, например, в (44.6) мы заменяем на каждом шаге норму матрицы, образованной измененными строками, на норму всей матрицы.

Если мы возьмем $A_0 = I$, то получим оценку для ошибки в вычисленном произведении $(n-1)$ -й приближенной матрицы отражения. Это дает

$$\begin{aligned} |fl_2(\bar{P}_{n-2} \bar{P}_{n-3} \dots \bar{P}_0) - P_{n-2} P_{n-3} \dots P_0|_E &< n^{3/2} x (1+x)^{n-2} = \\ &= (12,36) n^{3/2} 2^{-t} [1 + (12,36) n 2^{-t}]^{n-2}. \end{aligned} \quad (45.4)$$

Для подобных преобразований мы сразу же из (45.3) имеем результат с очевидной системой обозначений

$$\|\bar{A}_{n-1} - P_{n-2} P_{n-3} \dots P_0 A_0 P_0 \dots P_{n-2}\|_E \leq 2(n-1)(1+x)^{2n-4} x \|A_0\|_E. \quad (45.5)$$

Нет особых трудностей, типа рассмотренных в связи с плоскими вращениями, так как в нашем анализе не делается попытки установить специальные соотношения между частями последовательно преобразованных матриц. Очевидно, что умножения могли бы осуществляться в любом порядке, не влияя при этом на оценку.

Маловероятно, что множитель $n2^{-t}$ может быть улучшен для общей матрицы. Мы можем видеть это при рассмотрении следующего гипотетического примера. Предположим, что последовательность матриц получается таким путем. Нам даны n точных матриц отражения P_r ($r = 1, \dots, n$) и, начиная с A_0 , вычисляем n матриц $\bar{A}_r$ ($r = 1, \dots, n$). Матрица $\bar{A}_{r+1}$ получается из $\bar{A}_r$ точным вычислением $P_{r+1} \bar{A}_r P_{r+1}$ и последующим округлением ее элементов до t -разрядных чисел с плавающей запятой. Если мы напомним

$$\bar{A}_{r+1} = P_{r+1} \bar{A}_r P_{r+1} + G_{r+1}, \quad P_{r+1} \bar{A}_r P_{r+1} = B_{r+1}, \quad (45.6)$$

то очевидно, что оптимальная оценка для элементов G_{r+1} будет

$$|G_{r+1}|_{ij} \leq 2^{-t} |B_{r+1}|_{ij} \quad (45.7)$$

и, следовательно,

$$\|G_{r+1}\|_E \leq 2^{-t} \|B_{r+1}\|_E. \quad (45.8)$$

Наша оптимальная оценка для ошибки в окончательной матрице $\bar{A}_n$ дается неравенством

$$\|\bar{A}_n - P_n P_{n-1} \dots P_1 A_0 P_1 \dots P_{n-1} P_n\|_E \leqslant 2^{-t} [1 + (1 + 2^{-t}) + \dots + (1 + 2^{-t})^{n-1}] \|A_0\|_E, \quad (45.9)$$

и, следовательно, по существу она равна $n2^{-t} \|A_0\|_E$.

Если A_0 симметричная, то мы можем повсюду сохранить симметрию, но удобно так изменить вычисление, чтобы минимизировать объем работы. Мы обсудим это в главе 5, когда опишем соответствующие преобразования.

Неунитарные элементарные матрицы, аналогичные плоским вращениям

46. Большая численная устойчивость алгоритмов, основанных на унитарных преобразованиях, определяется тем, что евклидова и 2-норма инвариантны (если не учитывать ошибки округления), поэтому не может быть постепенного увеличения в целом элементов последовательно преобразованных матриц. Это важно, потому что на каждом шаге текущие ошибки округления в основном пропорциональны величине элементов преобразованных матриц.

Неунитарные элементарные матрицы не имеют свойств естественной устойчивости такого вида, но мы можем организовать вычисление так, чтобы дать им ограниченную форму устойчивости. Рассмотрим сначала проблему § 20. Там мы имели вектор $\bar{x}$ и делали его j -й элемент нулем с помощью левого умножения на матрицу, которая влияла лишь на i -й и j -й элементы, причем оба из них были заменены линейной комбинацией своих прежних значений. Теперь аналогичный эффект имеет левое умножение на матрицу M_{ji} (гл. 1, § 40 (vii)). Она оставляет без изменения все компоненты, за исключением j -й, которая определена равенством

$$x'_j = -m_{ji}\bar{x}_i + \bar{x}_j. \quad (46.1)$$

Мы можем сделать x'_j нулем, беря $m_{ji} = \bar{x}_j/\bar{x}_i$. Однако если мы сделаем так, значение m_{ji} может быть произвольно большим, возможно, даже бесконечным. Следовательно, $\|M_{ji}\|$ может быть произвольно большой и любая матрица A , которая умножается слева на M_{ji} , может иметь в своей j -й строке сколь угодно большие элементы. Из анализа, который мы дали, очевидно, что это нас не устраивает.

Мы можем повысить степень устойчивости следующим образом.

(i) Если $|\bar{x}_j| \leqslant |\bar{x}_i|$, то поступаем, как выше, и имеем $m_{ji} \leqslant 1$.

(ii) Если $|\bar{x}_j| > |\bar{x}_i|$, то сначала умножаем слева на I_{ij} (гл. 1, § 40 (i)), *переставляя* таким образом i -й и j -й элементы, и затем умножаем на соответствующую M_{ji} .

Итак, матрица, на которую мы умножаем слева, может быть представлена в виде $M_{ji}I_{ij}$, где I_{ij} есть или I_{ij} , или I , смотря по тому, была необходима *перестановка* или нет. Мы не рассмотрели случая, когда $\bar{x}_i = \bar{x}_j = 0$. Это не представляет никакой трудности; мы должны лишь взять

$$I_{ij} = I, \quad m_{ji} = 0, \quad (46.2)$$

так что $M_{ji}I'_{ij} = I$ в этом случае.

Матричное произведение $M_{ji}I_{ij}$ сравнимо по влиянию с вращением в плоскости (i, j) , и мы могли бы ожидать, что оно должно быть достаточно устойчиво. Будем обозначать это произведение через M'_{ji} , и простое вычисление показывает, что

$$\|M'_{ji}\|_2 \leq \left[\frac{1}{2} (3 + 5^{1/2}) \right]^{1/2}. \quad (46.3)$$

Аналогично алгоритмам, в которых мы используем последовательность вращений в плоскостях $(1, 2), (1, 3), \dots, (1, n); (2, 3), \dots, (2, n); \dots; (n-1, n)$, имеем алгоритмы, в которых используем совокупность матриц

$$M'_{21}, M'_{31}, \dots, M'_{n1}; M'_{32}, \dots, M'_{n2}; \dots; M'_{n, n-1}.$$

Мы не можем дать какой-либо общий анализ ошибок, приводящий к оценкам вида $n^k 2^{-t} \|A\|$; обычно для общих преобразований произвольных матриц самое лучшее, что можно достичь, так это оценки вида $2^n n^k 2^{-t} \|A\|$. Поэтому мы дадим лишь конкретный анализ в соответствующем месте в последующих главах.

Существует несколько алгоритмов, содержащих последовательности элементарных преобразований, в которых перестановки недопустимы, так как они могли бы уничтожить структуру нулей, полученных на предыдущих шагах. Чтобы выполнить эти алгоритмы, мы вынуждены пользоваться матрицами M_{ji} , а не M'_{ji} , и, следовательно, может возникнуть положение, в котором алгоритмы сколь угодно неустойчивы. Такие алгоритмы не имеют аналогов, основанных на плоских вращениях, так как их использование, аналогичное использованию M'_{ji} , неизбежно уничтожило бы структуру нулей.

Неунитарные элементарные матрицы, аналогичные матрицам отражения

47. Аналогичным путем, рассматривая проблему § 37, мы можем ввести неунитарные матрицы, которые аналогичны матрицам отражения. Задача состоит в том, чтобы по заданному вектору x определить матрицу M такую, чтобы Mx имела нули в компонентах от $(r+1)$ до n , тогда как компоненты с 1-й по r -ю не изменяются. Пусть рассматривается левое умножение x на N_r (гл. 1, § 40 (vi)).

Если мы напомним

$$x' = N_r x, \quad (47.1)$$

то

$$\left. \begin{aligned} x'_i &= x_i & (i=1, \dots, r), \\ x'_i &= x_i - n_{ir} x_r & (i=r+1, \dots, n), \end{aligned} \right\} \quad (47.2)$$

и беря

$$n_{ir} = \frac{x_i}{x_r} \quad (i=r+1, \dots, n), \quad (47.3)$$

мы видим, что условия удовлетворяются, но элементы N_r могут быть произвольно большими или даже бесконечными, и это будет приводить к численной неустойчивости. Устойчивость достигается следующим образом. Предположим, что максимум значения $|x_i|$ ($i=r, \dots, n$)

достигается для $i = r' \geq r$. Тогда вектор y определяется равенством

$$y = I_{r,r'} x, \quad (47.4)$$

т. е. получается из x перестановкой элементов x_r и $x_{r'}$. Следовательно,

$$|y_r| \geq |y_i| \quad (i = r + 1, \dots, n), \quad (47.5)$$

и если мы теперь выберем N_r так, чтобы $N_r y$ имел нулевые компоненты в позициях от $(r + 1)$ до n , то имеем

$$n_{ir} = \frac{y_i}{y_r} \quad (47.6)$$

и

$$|n_{ir}| \leq 1. \quad (47.7)$$

Итак, матрица N_r имеет элементы, которые по модулю ограничены единицей, и мы достигли в некоторой степени численной устойчивости. В действительности имеем

$$\|N_r\|_2 \leq \left\{ \frac{1}{2} [(n - r + 2) + (n - r)^{1/2} (n - r + 4)^{1/2}] \right\}^{1/2}. \quad (47.8)$$

Обозначим матричное произведение $N_r I_{r,r'}$ через N'_r . Если $x_i = 0$ ($i = r, \dots, n$), то берем $N'_r = I$. Очевидно, что N'_r может быть использована для достижения такого же эффекта, как при произведении $M'_{r+1,r}$, $M'_{r+2,r} \dots M'_{nr}$, хотя, начиная с того же x , мы, вообще говоря, не будем иметь

$$N'_r = M'_{r+1,r} M'_{r+2,r} \dots M'_{nr}. \quad (47.9)$$

Однако уравнение (47.9) будет удовлетворено, если не требуются перестановки. Процесс выбора в этом смысле наибольшего элемента из заданной совокупности обычно называется выбором главного элемента.

Аналогично алгоритмам, в которых мы используем совокупность матриц отражения $(I - 2w_r w_r^T)$ с w_r , имеющими нулевые компоненты в позициях с 1-й по $(r - 1)$ -ю, мы имеем алгоритмы, в которых используем матрицы $N'_1, N'_2, \dots, N'_{n-1}$. Мы опять, как правило, не сможем дать общий анализ ошибок, приводящий к оценкам, аналогичным тем, которые дали для преобразований, основанных на матрицах отражения. Будем называть M'_{ji} и N'_r *устойчивыми элементарными матрицами*.

Как и раньше, найдем, что существует несколько алгоритмов, основанных на элементарных подобных преобразованиях, в которых недопустимо использование перестановок на каждом шаге, так как это могло бы разрушить важную структуру нулей, полученных предыдущими преобразованиями. Некоторые алгоритмы могут быть неустойчивыми и нуждаются в тщательном рассмотрении. Они не имеют аналогов, основанных на матрицах отражения, так как их использование неизбежно уничтожало бы структуры нулей таким же образом, как и использование перестановок.

Повсюду в этой книге мы неоднократно будем сталкиваться с необходимостью выбора между использованием элементарных унитарных матриц и элементарных устойчивых матриц. Вообще говоря, элементарные устойчивые преобразования будут несколько проще, но, с другой стороны, мы не сможем дать для них столь хорошей гарантии численной устойчивости.

Левые умножения на последовательность неунитарных матриц

48. Существует одно преобразование, использующее матрицы типа N'_r , для которого мы сможем дать достаточно удовлетворительный анализ ошибок. Это левое умножение матрицы A_0 на последовательность матриц $\bar{N}'_1, \bar{N}'_2, \dots, \bar{N}'_{n-1}$. Мы запишем

$$\bar{A}_r = fl(\bar{N}'_r \bar{A}_{r-1}) \equiv \bar{N}'_r \bar{A}_{r-1} + H_r, \quad (48.1)$$

так что H_r есть матрица ошибок, сделанных при действительном умножении. Левое умножение на I_r , r' лишь переставляет строки r и r' и поэтому не влечет за собой ошибок округления. В последующем левом умножении на $\bar{N}'_r$ строки с 1-й до r -й не изменяются, так что H_r в этих строках нулевая. Объединяя уравнения (48.1) для r от 1 до $n-1$, мы доказываем аналогично доказательству § 15, что

$$\begin{aligned} \bar{A}_{n-1} = \bar{N}'_{n-1} \bar{N}'_{n-2} \dots \bar{N}'_1 [A_0 + (\bar{N}'_1)^{-1} H_1 + (\bar{N}'_1)^{-1} (\bar{N}'_2)^{-1} H_2 + \dots \\ \dots + (\bar{N}'_1)^{-1} \dots (\bar{N}'_{n-1})^{-1} H_{n-1}]. \end{aligned} \quad (48.2)$$

Поэтому матрица $\bar{A}_{n-1}$ в точности эквивалентна матрице в квадратных скобках правой части (48.2).

Теперь покажем, что

$$(\bar{N}'_1)^{-1} (\bar{N}'_2)^{-1} \dots (\bar{N}'_r)^{-1} H_r = I_{1,1} I_{2,2'} \dots I_{r,r'} H_r. \quad (48.3)$$

Имеем

$$(\bar{N}'_r)^{-1} H_r = (\bar{N}_r I_{r,r'})^{-1} H_r = I_{r,r'} \bar{N}_r^{-1} H_r. \quad (48.4)$$

Здесь $\bar{N}_r^{-1}$ такова же, как матрица $\bar{N}_r$, но с измененными знаками элементов, лежащих ниже диагонали. Следовательно,

$$\bar{N}_r^{-1} H_r = H_r, \quad (48.5)$$

так как левое умножение на $\bar{N}_r^{-1}$ приводит лишь к прибавлению кратного нулевой r -й строки к строкам от $(r+1)$ до n . Поэтому

$$(\bar{N}'_r)^{-1} H_r = I_{r,r'} H_r, \quad (48.6)$$

и так как $r' \geq r$, матрица $I_{r,r'} H_r$ нулевая в строках от 1 до $(r-1)$. Продолжая это доказательство, получаем (48.3). Матрица в правой части (48.3) есть просто H_r с переставленными строками. Следовательно, уравнение (48.2) показывает, что $\bar{A}_{n-1}$ в точности эквивалентна A_0 с возмущением

$$I_{1,1} H_1 + I_{1,1} I_{2,2'} H_2 + \dots + I_{1,1} I_{2,2'} \dots I_{n-1,(n-1)} H_{n-1}. \quad (48.7)$$

Итак, каждая совокупность ошибок H_r эквивалентна по своему влиянию возмущению исходной матрицы, состоящему из матриц H_r с переставленными строками. Никакого увеличения этих ошибок не происходит, если они включены в исходную матрицу. (Заметим, что этот результат верен и тогда, когда не используются перестановки.) Существует, конечно, аналогичный результат для правого умножения на матрицы вида M'_r , где

$$M'_r = I_{r,r'} M_r, \quad (48.8)$$

и правое умножение на I_r , r' здесь осуществляет перестановку столбцов.

Из этого замечательного результата непосредственно вытекает, что опасность неустойчивости возникает лишь из того, что сами матрицы H_r

могут быть большими. Так как ошибки, сделанные на любом шаге, вообще говоря, пропорциональны величине текущих элементов преобразованной матрицы, то мы видим чрезвычайную важность обеспечения того, чтобы элементы A , были как можно меньшими. По этой причине перестановки существенны.

Читателю следует проверить, что соответствующий анализ для преобразований с матрицами M_{ji} не дает столь удовлетворительного результата. Эквивалентные возмущения в исходной матрице имеют значительно более сложную структуру.

Априорные оценки ошибок

49. Мы видели, что для некоторых алгоритмов, основанных на унитарных преобразованиях, можно предполагать возможность получения априорной оценки для возмущения в исходной матрице, которое эквивалентно ошибкам, сделанным при вычислении. Это не дает нам, вообще говоря, априорных оценок для ошибок в тех величинах, которые мы хотим вычислить, хотя в случае подобных преобразований симметричных матриц можно воспользоваться результатами главы 2, если требуемые величины являются собственными значениями. Но даже и в этом очень благоприятном случае мы не можем получить априорные оценки для ошибок в собственных векторах. Однако априорные оценки для эквивалентных возмущений в исходной матрице имеют большое значение при сравнении алгоритмов.

Мы найдем, что наиболее эффективные оценки, которые сможем получить для эквивалентного возмущения, имеют вид

$$\|F\| \leq n^k 2^{-t} \|A\|, \quad (49.1)$$

где k — какая-то константа, обычно меньшая трех. Такие оценки, как правило, можно получить лишь для алгоритмов, основанных на унитарных преобразованиях. Для устойчивых элементарных неунитарных преобразований обычно достаточно трудно получить даже оценки вида

$$\|F\| \leq n^k 2^n 2^{-t} \|A\|, \quad (49.2)$$

хотя для конкретных методов существуют основания верить в то, что множитель 2^n может быть достигнут лишь при крайне неблагоприятных условиях. Для неустойчивых неунитарных преобразований даже оценки вида (49.2) нельзя получить. Действительно, как правило, будут существовать обстоятельства, вообще не имеющие отношения к плохой обусловленности исходной задачи, при которых такие преобразования будут полностью нарушаться из-за деления на нуль.

Отклонение от нормальности

50. Для того чтобы получить оценки ошибок, получающихся при использовании данного алгоритма, нам нужны не только оценки для эквивалентных возмущений в исходной матрице, но и оценки чувствительности требуемых величин. Естественно спросить, существует ли какой-либо простой путь оценки фактора чувствительности в случае проблемы собственных значений. Как мы знаем из главы 2, (31.2), если A нормальная, то собственные значения $(A + F)$ лежат в кругах

$$|\lambda - \lambda_i| \leq \|F\|_2 \leq \|F\|_E. \quad (50.1)$$

Поэтому мы могли бы ожидать, что если A «почти нормальная матрица», то ее собственные значения должны быть мало чувствительны. Если A — нормальная, то (гл. 1, § 48) существует унитарная матрица R такая, что

$$R^H A R = \text{diag}(\lambda_i), \quad (50.2)$$

тогда как (гл. 1, § 47) для произвольной матрицы существует унитарная R такая, что

$$R^H A R = \text{diag}(\lambda_i) + T = D + T, \quad (50.3)$$

где T — строго верхняя треугольная, т. е.

$$t_{ij} = 0 \quad (i \geq j). \quad (50.4)$$

Было бы разумно рассматривать $\|T\|/\|A\|$ как меру отклонения A от нормальности.

Хенричи (1962) использовал соотношение (50.3) для получения оценок влияния возмущений в A на ее собственные значения следующим образом.

Пусть G определена соотношением

$$R^H (A + F) R = D + T + G, \quad (50.5)$$

так что

$$G = R^H F R, \quad \|G\|_2 = \|F\|_2, \quad \|G\|_E = \|F\|_E. \quad (50.6)$$

Если λ есть некоторое собственное значение матрицы $(A + F)$, то $(D - \lambda I) + T + G$ вырождена. Мы различаем два случая:

- (i) $\lambda = \lambda_i$ для некоторого i ,
- (ii) $\lambda \neq \lambda_i$ для любого i , при этом $(D - \lambda I)$ и $(D + T - \lambda I)$ не вырождены.

Итак,

$$0 = \det[D + T - \lambda I + G] = \det[D + T - \lambda I] \det[I + (D + T - \lambda I)^{-1} G], \quad (50.7)$$

откуда

$$0 = \det[I + (D + T - \lambda I)^{-1} G]. \quad (50.8)$$

Уравнение (50.8) означает, что

$$\begin{aligned} 1 &\leq \| (D + T - \lambda I)^{-1} G \|_2 \leq \\ &\leq \| (D + T - \lambda I)^{-1} \|_2 \| G \|_2 = \\ &= \| (I + KT)^{-1} K \|_2 \| G \|_2 = \\ &= \| [I - KT + (KT)^2 - \dots + (-1)^{n-1} (KT)^{n-1}] K \|_2 \| G \|_2, \end{aligned} \quad (50.9)$$

где

$$K = (D - \lambda I)^{-1} = \text{diag}[(\lambda_i - \lambda)^{-1}], \quad (50.10)$$

$(KT)^r = 0$ ($r \geq n$) в силу того, что KT — строго верхняя треугольная. Полагая

$$\|K\|_2 = a, \quad \|T\|_2 = b, \quad \|T\|_E = c, \quad (50.11)$$

имеем

$$1 \leq (1 + ab + a^2 b^2 + \dots + a^{n-1} b^{n-1}) a \|G\|_2 \leq \quad (50.12)$$

$$\leq (1 + ac + a^2 c^2 + \dots + a^{n-1} c^{n-1}) a \|G\|_E. \quad (50.13)$$

Заметим, что этот результат верен и тогда, когда мы заменяем c на любую большую величину.

Далее

$$a = \|K\|_2 = 1/\min |\lambda_i - \lambda|, \quad (50.14)$$

и, следовательно, (50.13) показывает, что существует λ_i такое, что

$$\frac{|\lambda_i - \lambda|}{1 + ac + a^2c^2 + \dots + a^{n-1}c^{n-1}} \leq \|G\|_2 = \|F\|_2 \leq \|F\|_E. \quad (50.15)$$

Этот же результат имеет место и тогда, когда c заменяется на b .

Так как в случае (i) очевидно, что соотношение (50.15) верно, оно верно и во всех случаях. Если A — нормальная, то $c = 0$ приводит к (50.4).

Простые примеры

51. Этот результат свободен от ограничений, которые мы обсудили в гл. 2, § 26. Если мы возьмем, например,

$$A = \begin{bmatrix} 1 & \varepsilon & & \\ & 1 & \varepsilon & \\ & & \ddots & \\ & & & 1 & \varepsilon \\ & & & & 1 \end{bmatrix}, \quad \|F\|_E = \eta, \quad (51.1)$$

то $b = \varepsilon$, и положив $|1 - \lambda| = \xi$, имеем

$$\frac{\xi}{1 + (\varepsilon/\xi) + \dots + (\varepsilon/\xi)^{n-1}} \leq \eta. \quad (51.2)$$

Если $n = 10$, $\varepsilon = 10^{-10}$, $\eta = 10^{-8}$, то максимум значения ξ есть решение уравнения

$$\xi = 10^{-8} [1 + (10^{-10}/\xi) + (10^{-10}/\xi)^2 + \dots + (10^{-10}/\xi)^9], \quad (51.3)$$

и очевидно, что он лежит между 10^{-8} и $10^{-8} (1 + 10^{-2})$.

Чтобы сделать доступным использование (50.15), нам нужна верхняя оценка для c , которая могла бы быть вычислена достаточно легко. Для того чтобы A была нормальной, необходимо и достаточно, чтобы $(A^H A - A A^H)$ была нулевой. Поэтому мы могли бы с достаточным основанием ожидать, что c связано с $(A^H A - A A^H)$. Хенричи (1962) показал в действительности, что

$$c \leq \left(\frac{n^3 - n}{12} \right)^{1/4} \|A^H A - A A^H\|_E^{1/2}, \quad (51.4)$$

а это требует выполнения двух матричных умножений. Так как для матричного умножения необходимо n^3 умножений, то объем работы далеко не мал.

К сожалению, матрица может иметь значительное отклонение от нормальности, будучи просто плохо обусловленной. Рассмотрим, например, матрицу

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}. \quad (51.5)$$

Здесь мы имеем $b = c = 2$ и, следовательно, если λ есть собственное значение $(A + F)$, то наш результат показывает, что или

$$\frac{|\lambda - 1|}{1 + 2/|\lambda - 1|} \leq \|F\|_E = \eta, \quad (51.6)$$

или

$$\frac{|\lambda - 3|}{1 + 2/|\lambda - 3|} \leq \|F\|_E = \eta. \quad (51.7)$$

Это означает, что или

$$\left. \begin{aligned} |\lambda - 1| &\leq \frac{1}{2} [\eta + (\eta^2 + 8\eta)^{1/2}], \\ |\lambda - 3| &\leq \frac{1}{2} [\eta + (\eta^2 + 8\eta)^{1/2}], \end{aligned} \right\} \quad (51.8)$$

и радиусы этих кругов больше чем $(2\eta)^{1/2}$. Левые и правые собственные векторы этой матрицы таковы:

$$\left. \begin{aligned} \lambda_1 = 3; \quad y_1^T &= [0, 1] \quad \text{и} \quad x_1^T = [1, 1]; \\ \lambda_2 = 1; \quad y_2^T &= [1, -1] \quad \text{и} \quad x_2^T = [1, 0]. \end{aligned} \right\} \quad (51.9)$$

Итак (гл. 2, § 8), $|s_1| = |s_2| = 2^{-1/2}$. Если k есть спектральное число обусловленности, то из соотношения (31.8) главы 2 следует, что $k = |s_1^{-1}| + |s_2^{-1}| \leq 2^{3/2}$ и, согласно (30.7) из главы 2, собственные значения $(A + F)$ лежат в кругах

$$\left. \begin{aligned} |\lambda - 1| &\leq 2^{3/2}\eta, \\ |\lambda - 3| &\leq 2^{3/2}\eta. \end{aligned} \right\} \quad (51.10)$$

Очевидно, что когда η мало, результаты (51.8) очень слабы.

Апостериорные оценки

52. В настоящее время, по-видимому, нет каких-либо пригодных оценок для возмущений собственных значений ненормальных матриц общего вида. Даже довольно слабый результат предыдущего параграфа требует $2n^3$ умножений для того, чтобы получить оценку отклонения от нормальности.

В настоящее время некоторые алгоритмы, которые мы опишем, требуют лишь порядка n^3 умножений для решения полной проблемы собственных значений. Не будет неожиданностью, что из приближенно вычисленной собственной системы мы можем получить значительно более точные оценки для собственных значений. Оценки ошибок, полученные из вычисленной полной или частичной собственной системы, называются апостериорными оценками. В подавляющем большинстве случаев для ненормальных матриц, по-видимому, нужно полагаться на апостериорные оценки.

Апостериорные оценки для нормальных матриц

53. Предположим, что мы имеем приближенное собственное значение μ и соответствующий собственный вектор v матрицы A . Тогда рассмотрим вектор η , определенный равенством

$$(A - \mu I) v = A v - \mu v = \eta. \quad (53.1)$$

Обычно он называется *вектором-невязкой*, соответствующим μ и v . Если μ и v точные, то η должен быть нулевым. Следовательно, если η мал, мы могли бы ожидать, что μ будет хорошим приближением к собственному значению.

Если A имеет линейные делители, то существует H такая, что

$$H \operatorname{diag}(\lambda_i) H^{-1} = A.$$

Итак, из (53.1)

$$H \operatorname{diag}(\lambda_i - \mu) H^{-1} v = \eta. \quad (53.2)$$

Возможны два случая:

- (i) $\lambda_i = \mu$ для некоторого i ;
- (ii) $\lambda_i \neq \mu$ для любого i , при этом (53.2) дает

$$v = H \operatorname{diag}[(\lambda_i - \mu)^{-1}] H^{-1} \eta, \quad (53.3)$$

и, следовательно,

$$\|v\| \leq \|H\| \|H^{-1}\| \|\operatorname{diag}[(\lambda_i - \mu)^{-1}]\| \|\eta\| = \max_i \frac{1}{|\mu - \lambda_i|} \|H\| \|H^{-1}\| \|\eta\| \quad (53.4)$$

для любой нормы такой, что норма диагональной матрицы равна модулю наибольшего ее элемента.

Таким образом, используя 2-норму,

$$\min |\mu - \lambda_i| \leq k \|\eta\|_2 / \|v\|_2, \quad (53.5)$$

где k такое же, как в § 51. Соотношение (53.5), очевидно, удовлетворяется в случае (i), и, следовательно, мы можем заключить, что всегда существует по крайней мере одно собственное значение матрицы A в круге с центром μ и радиусом $k \|\eta\|_2 / \|v\|_2$. Очевидно, что не теряется никакой общности в предположении, что $\|v\|_2 = 1$, и мы будем делать его в дальнейшем. Так как, вообще говоря, мы не имеем априорной оценки k , то этот результат обычно практически не очень важен, если только у нас нет больше информации о собственной системе. Однако если A — нормальная, то мы знаем, что $k = 1$, и тогда наш результат показывает, что существует по крайней мере одно собственное значение в круге с центром μ и радиусом $\|\eta\|_2$.

Существует другой путь рассмотрения уравнения (53.1), который связывает этот результат с результатом главы 2. Уравнение (53.1) означает, что

$$(A - \eta v^H) v = A v - \eta (v^H v) = A v - \eta = \mu v, \quad (53.6)$$

и, следовательно, μ и v будут *точным* собственным значением и собственным вектором матрицы $(A - \eta v^H)$. Из гл. 2, § 30 мы знаем, что все собственные значения $(A - \eta v^H)$ лежат в кругах

$$|\lambda - \lambda_i| \leq k \|\eta v^H\|_2 = k \|\eta\|_2 \quad (i = 1, \dots, n), \quad (53.7)$$

и это приводит к тому же самому результату, который мы только что получили.

Отношение Релея

54. Если дан произвольный вектор v , то мы можем выбрать значение μ так, чтобы $\|Av - \mu v\|_2$ была минимальной. Обозначив $k = v^H Av$ и предполагая, что $\|v\|_2 = 1$, имеем

$$\|Av - \mu v\|_2^2 = (Av - \mu v)^H (Av - \mu v) = v^H A^H Av - k k^H + (\mu^H - k^H)(\mu - k). \quad (54.1)$$

Очевидно, что минимум достигается для $\mu = k = v^H Av$. Эта величина называется *отношением Релея*, соответствующим v . Мы обозначим его через μ_R . Термин «отношение» используется потому, что если v ненормирован, то соответствующее значение есть $v^H Av / v^H v$. Очевидно, что если

$$Av - \mu_R v = \eta_R, \quad (54.2)$$

то

$$v^H \eta_R = 0. \quad (54.3)$$

Если для некоторого μ мы имеем

$$Av - \mu v = \eta \quad (\|\eta\|_2 = \varepsilon), \quad (54.4)$$

то

$$Av - \mu_R v = \eta_R \quad (\|\eta_R\|_2 \leq \varepsilon). \quad (54.5)$$

Круг с центром μ_R имеет радиус, который по крайней мере настолько же мал, как и радиус, соответствующий любому другому μ . Отношение Релея определено для любой матрицы, но оно имеет особое значение для нормальных матриц, что мы сейчас покажем. Предположим, что мы имеем

$$Av - \mu_R v = \eta \quad (\|v\|_2 = 1, \|\eta\|_2 = \varepsilon), \quad (54.6)$$

и знаем, что $n - 1$ из собственных значений удовлетворяют соотношению

$$|\lambda_i - \mu_R| \geq a, \quad (54.7)$$

где a — некоторая константа. Можем считать соответствующие собственные значения равными $\lambda_2, \lambda_3, \dots, \lambda_n$. Теперь, если A нормальная, то она, конечно, имеет ортогональную систему собственных векторов x_i ($i = 1, \dots, n$), и мы можем написать

$$v = \sum_1^n \alpha_i x_i, \quad 1 = \sum_1^n |\alpha_i|^2. \quad (54.8)$$

Уравнение (54.6) поэтому дает

$$\eta = \sum_1^n \alpha_i (\lambda_i - \mu_R) x_i,$$

и, следовательно,

$$\varepsilon^2 = \sum_1^n |\alpha_i|^2 |\lambda_i - \mu_R|^2 \geq \sum_2^n |\alpha_i|^2 |\lambda_i - \mu_R|^2 \geq \quad (54.9)$$

$$\geq a^2 \sum_2^n |\alpha_i|^2 \quad \text{из (54.7)} \quad (54.10)$$

Отсюда выводим, что

$$|\alpha_1|^2 = 1 - \sum_2^n |\alpha_i|^2 \geq 1 - \varepsilon^2 / a^2. \quad (54.11)$$

Если мы определим θ соотношением

$$\alpha_1 = |\alpha_1| e^{i\theta}, \quad (54.12)$$

то

$$ve^{-i\theta} = |\alpha_1| x_1 + \sum_2^n \alpha_i e^{-i\theta} x_i \quad (54.13)$$

и, следовательно,

$$\|ve^{-i\theta} - x_1\|_2^2 = (|\alpha_1| - 1)^2 + \sum_2^n |\alpha_i|^2 \leq \frac{\varepsilon^4/a^4}{(1 + |\alpha_1|^2)} + \frac{\varepsilon^2}{a^2} \leq \frac{\varepsilon^4}{a^4} + \frac{\varepsilon^2}{a^2}. \quad (54.14)$$

Из (54.14) видим, что если $a \gg \varepsilon$, то $ve^{-i\theta}$ есть хорошее приближение к x_1 . Постоянный множитель $e^{-i\theta}$ имеет небольшое значение. По мере того как a уменьшается, результат постепенно ослабевает до тех пор, пока a не будет величиной порядка ε , когда он больше уже непригоден. Заметим, что результаты, полученные до сих пор, не предполагают, что μ_R есть отношение Релея.

Ошибка в отношении Релея

55. Возвращаясь теперь к μ_R , имеем из его определения

$$\mu_R = v^H A v = \sum_1^n \lambda_i |\alpha_i|^2. \quad (55.1)$$

Следовательно,

$$\mu_R \sum_1^n |\alpha_i|^2 = \sum_1^n \lambda_i |\alpha_i|^2,$$

откуда

$$(\mu_R - \lambda_1) |\alpha_1|^2 = \sum_2^n (\lambda_i - \mu_R) |\alpha_i|^2 = \sum_2^n |\lambda_i - \mu_R|^2 |\alpha_i|^2 / (\lambda_i - \mu_R)^H. \quad (55.2)$$

Из (54.11), (54.7) и (54.9) находим

$$\begin{aligned} \left(1 - \frac{\varepsilon^2}{a^2}\right) |\mu_R - \lambda_1| &\leq \sum_2^n |\lambda_i - \mu_R|^2 |\alpha_i|^2 / |\lambda_i - \mu_R| \leq \\ &\leq \sum_2^n |\lambda_i - \mu_R|^2 |\alpha_i|^2 / a \leq \varepsilon^2 / a, \\ |\mu_R - \lambda_1| &\leq \frac{\varepsilon^2}{a} \left/ \left(1 - \frac{\varepsilon^2}{a^2}\right) \right. \end{aligned} \quad (55.3)$$

Соотношение (55.3) показывает, что если $a \gg \varepsilon$, отношение Релея имеет ошибку порядка ε^2 .

По мере того как a уменьшается, результат постепенно ослабевает, до тех пор, пока a не будет величиной того же порядка, что и ε , когда он уже не сильнее результата для произвольного μ такого, что $\|Av - \mu v\|_2 = \varepsilon$.

Подобное ухудшение оценок ошибок в отношении Релея и векторе неизбежно, когда a порядка ε , что иллюстрируется следующим примером. Рассмотрим матрицу A и предполагаемые собственное значение и собственный вектор, заданные так:

$$A = \begin{bmatrix} b & \varepsilon \\ \varepsilon & b \end{bmatrix}, \quad \mu_R = b, \quad v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (55.4)$$

Имеем

$$Av - \mu_R v = \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix}, \quad (55.5)$$

и, следовательно, $\| \eta \|_2 = \varepsilon$ для нормированного вектора v . Собственные значения будут $b \pm \varepsilon$, так что μ_R имеет ошибку порядка ε . Собственные векторы суть

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad (55.6)$$

и это показывает, что v ни в каком смысле не будет приближенным собственным вектором.

Эрмитовы матрицы

56. Единственные нормальные матрицы, с которыми мы будем часто иметь дело, это эрмитовы матрицы, обычно вещественные и симметричные. Решение собственной проблемы для комплексной эрмитовой матрицы может быть сведено к решению проблемы для вещественной симметричной матрицы следующим способом. Предположим, что

$$\left. \begin{aligned} A &= B + iC & (B \text{ и } C \text{ действительные}), \\ B^T &= B, & C^T = -C. \end{aligned} \right\} \quad (56.1)$$

Тогда, если λ_i — собственное значение A (обязательно вещественное) и $(u_i + iv_i)$ — соответствующий собственный вектор, то мы имеем

$$(B + iC)(u_i + iv_i) = \lambda_i(u_i + iv_i),$$

т. е.

$$\left. \begin{aligned} Bu_i - Cv_i &= \lambda_i u_i, \\ Cu_i + Bv_i &= \lambda_i v_i. \end{aligned} \right\} \quad (56.2)$$

Следовательно,

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \lambda_i u_i \\ \lambda_i v_i \end{bmatrix}, \quad \begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} v_i \\ -u_i \end{bmatrix} = \begin{bmatrix} \lambda_i v_i \\ -\lambda_i u_i \end{bmatrix}, \quad (56.3)$$

и, так как

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} v_i \\ -u_i \end{bmatrix} \quad (56.4)$$

ортогональны, расширенная вещественная матрица имеет λ_i двукратным собственным значением. Очевидно, что расширенная матрица будет вещественная и симметричная и ее собственные значения будут

$$\lambda_1, \lambda_1, \lambda_2, \lambda_2, \dots, \lambda_n, \lambda_n.$$

Большая часть лучших алгоритмов определения собственной системы вещественных симметричных матриц приводит к вещественным приближениям собственных значений и собственных векторов. Поэтому круги о которых мы упоминали, становятся интервалами на вещественной оси.

Заметим, что если μ_R есть отношение Релея, соответствующее данному v , и

$$Av - \mu_R v = \eta, \quad (56.5)$$

то, так как $\eta^T v = 0$, мы имеем

$$(A - \eta v^T - v \eta^T) v = \mu_R v, \quad (56.6)$$

так что v и μ_R являются точными собственным вектором и собственным значением для матрицы $(A - \eta v^T - v \eta^T)$. Если A симметричная, то эта матрица также симметричная, и мы имеем дело с симметричным возмущением симметричной матрицы. Легко проверить, что

$$\|\eta v^T + v \eta^T\|_2 = \|\eta\|_2, \quad (56.7)$$

если $\|v\|_2 = 1$ и $\eta^T v = 0$.

Если мы получим n приближенных собственных значений и собственных векторов и интервалы разделены, то мы знаем, что в каждом из них существует точно одно собственное значение. Результат последнего параграфа даст нам возможность локализовать собственные значения очень точно, если только они неплохо отделены. Таким образом, если для матрицы четвертого порядка имеем отношения Релея μ_i и радиусы η_i , определенные так:

i	1	2	3	4
μ_i	0,1	0,1001	0,2	0,3
$\ \eta_i\ _2$	10^{-8}	$2 \cdot 10^{-8}$	$2 \cdot 10^{-8}$	$3 \cdot 10^{-8}$

(56.8)

то мы сразу же видим, что существует по крайней мере одно собственное значение в каждом из интервалов

$$\left. \begin{aligned} |\lambda - 0,1| &\leq 10^{-8}, & |\lambda - 0,1001| &\leq 2 \cdot 10^{-8}, \\ |\lambda - 0,2| &\leq 2 \cdot 10^{-8}, & |\lambda - 0,3| &\leq 3 \cdot 10^{-8}. \end{aligned} \right\} \quad (56.9)$$

Очевидно, что эти интервалы разделены и, следовательно, каждый содержит точно одно собственное значение.

Теперь мы проиллюстрируем использование результата § 55, получив очень точную оценку для собственного значения λ_2 во втором интервале. Очевидно, что

$$\left. \begin{aligned} |\mu_2 - \lambda_1| &\geq |0,1001 - 0,1 - 10^{-8}| = 10^{-4}(0,9999), \\ |\mu_2 - \lambda_3| &\geq |0,1001 - 0,2 + 2 \cdot 10^{-8}| > 0,0998, \\ \mu_2 - \lambda_4 &\geq |0,1001 - 0,3 + 3 \cdot 10^{-8}| > 0,1998. \end{aligned} \right\} \quad (56.10)$$

Следовательно, мы можем взять $a = 10^{-4}(0,9999)$, и (55.3) дает

$$|\mu_2 - \lambda_2| \leq \frac{4 \cdot 10^{-16}}{10^{-4}(0,9999)} \left/ \left\{ 1 - \left[\frac{2 \cdot 10^{-8}}{10^{-4}(0,9999)} \right]^2 \right\} \right. = (4,0004) 10^{-12}.$$

Несмотря на то, что λ_1 и λ_2 могут считаться совсем близкими, отношение Релея очень точно. Частные μ_3 и μ_4 даже более точные. На практике мы найдем, что отношение Релея настолько точно (за исключением патологически близких собственных значений), что оно должно вычисляться с большой осторожностью, если нужно реализовать полностью его возможности.

Патологически близкие собственные значения

57. Если интервалы, в которых лежат отношения Релея, не разделены, то положение значительно более неясно, и без дальнейшего исследования мы не можем быть уверены, что все собственные значения согласуются с интервалами. Мы можем проиллюстрировать трудность рассмотрением предельного случая. Предположим, что собственная проблема матрицы четвертого порядка с хорошо разделенными собственными значениями решена, используя четыре разных алгоритма. Мы обозначим приближения к собственному значению λ_i и собственному вектору x_i , полученные по четырем методам, через μ_i и v_i ($i = 1, \dots, 4$). Хотя все эти приближения к одному и тому же собственному значению и собственному вектору, они, вообще говоря, будут неодинаковы; в действительности обычно v_i строго линейно независимы. Если даны эти четыре пары μ_i и v_i , мы получим четыре перекрывающихся интервала, и (предполагая, что алгоритмы были устойчивые) тем не менее в объединении этих интервалов будет лишь одно собственное значение. Очевидно, что нам нужно будет много больше, чем простая математическая независимость векторов, если мы хотим сделать полезные и строгие выводы.

Следующие соображения проливают свет на эту проблему. Предположим, что ортогональные нормированные векторы v_i и μ_i ($i = 1, \dots, r$) таковы, что

$$A v_i - \mu_i v_i = \eta_i, \quad \|\eta_i\|_2 = \varepsilon_i. \quad (57.1)$$

Существует последовательность нормированных векторов $v_{r+1}, \dots, v_n$ таких, что матрица V , имеющая v_i ($i = 1, 2, \dots, n$) своими столбцами, ортогональна. Рассмотрим теперь матрицу B , определенную равенством

$$B = V^T A V. \quad (57.2)$$

Очевидно, что B — симметричная, и из (57.1) имеем

$$\begin{aligned} b_{ij} &= v_i^T A v_j & (\text{все } i, j), \\ b_{ij} &= v_i^T (\mu_j v_j + \eta_j) & (j = 1, \dots, r). \end{aligned} \quad (57.3)$$

Учитывая симметрию, можем написать

$$B = \left[\begin{array}{ccc|c} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_r \\ \hline 0 & & & X \end{array} \right] + \left[\begin{array}{c|c} P & Q \\ \hline Q^T & O \end{array} \right], \quad (57.4)$$

где X — симметричная и

$$[P \mid Q]_{ij} = v_i^T \eta_j \quad (i = 1, \dots, r; j = 1, \dots, n). \quad (57.5)$$

Далее, B имеет собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$ матрицы A и если $\mu_{r+1}, \mu_{r+2}, \dots, \mu_n$ — собственные значения матрицы X , то по теореме

Виландта — Гофмана (гл. 2, § 48) для некоторой нумерации λ_i имеем

$$\begin{aligned} \sum_1^r (\lambda_i - \mu_i)^2 &\leq \sum_1^r (\lambda_i - \mu_i)^2 \leq \|P\|_E^2 + \|Q\|_E^2 + \|Q^T\|_E^2 \leq 2(\|P\|_E^2 + \|Q\|_E^2) = \\ &= 2 \sum_1^r \|\eta_i\|^2 = 2 \sum_1^r \varepsilon_i^2 = k^2. \end{aligned} \quad (57.6)$$

Итак, несомненно существует r собственных значений в объединении r интервалов $\mu_i - k \leq \lambda \leq \mu_i + k$. Заметим, что μ_i не обязательно должны быть отношениями Релея. Например, рассмотрим диагональную матрицу D порядка n с элементами

$$d_{ii} = a \quad (i = 1, 2, \dots, n-1), \quad d_{nn} = a + n^{1/2}\varepsilon. \quad (57.7)$$

Если мы возьмем

$$v = (\pm n^{-1/2}, \pm n^{-1/2}, \dots, \pm n^{-1/2}), \quad \mu = a, \quad (57.8)$$

то имеем

$$\eta = (0, 0, \dots, \pm \varepsilon) \quad (57.9)$$

и, следовательно, какое бы ни было распределение знаков, $\|\eta\|_2 = \varepsilon$. Хорошо известно, что если n есть степень числа 2, то существует n ортогональных векторов типа (57.8). Все соответствующие интервалы имеют один и тот же центр a и очевидно, что наименьший интервал, который будет покрывать все n собственных значений, есть $(a - n^{1/2}\varepsilon)$ до $(a + n^{1/2}\varepsilon)$. Отношение Релея, соответствующее каждому из v , равно $(a + n^{1/2}\varepsilon)$, так что мы не получим значительно лучшего результата.

Предположим теперь, что даны

$$A v_1 - \mu_1 v_1 = \eta_1, \quad (57.10)$$

$$A v_2 - \mu_2 v_2 = \eta_2, \quad (57.11)$$

где $\|\eta_1\|_2, \|\eta_2\|_2, |\mu_1 - \mu_2|$ — малые величины и

$$v_1^T v_2 = \cos \theta. \quad (57.12)$$

Мы можем написать

$$v_2 = v_1 \cos \theta + w_1 \sin \theta, \quad (57.13)$$

где w_1 — единичный вектор в подпространстве, натянутом на v_1 и v_2 , ортогональный к v_1 . Умножая (57.10) на $\cos \theta$ и вычитая из (57.11), получаем

$$A w_1 \sin \theta - (\mu_2 - \mu_1) v_1 \cos \theta - \mu_2 w_1 \sin \theta = \eta_2 - \eta_1 \cos \theta,$$

откуда

$$A w_1 - \mu_2 w_1 = \eta_3, \quad (57.14)$$

где

$$\eta_3 = \frac{\eta_2 - \eta_1 \cos \theta + (\mu_2 - \mu_1) v_1 \cos \theta}{\sin \theta}, \quad (57.15)$$

$$\|\eta_3\|_2 \leq \frac{\|\eta_2\|_2 + \|\eta_1\|_2 + |\mu_2 - \mu_1|}{|\sin \theta|}. \quad (57.16)$$

При условии, что $\sin \theta$ не мал, (57.14) и (57.16) теперь показывают, что ортогональные векторы v_1 и w_1 имеют малые невязки, и мы можем использовать наш предыдущий результат. Если θ уменьшается, этот результат постепенно ослабевает.

Ненормальные матрицы

58. Для нормальных матриц мы часто можем получить очень точные и строгие оценки собственных значений из относительно неточной информации, относящейся лишь к части собственной системы. Для ненормальных матриц это возможно редко. Соотношение (53.5) имеет значение лишь тогда, когда мы имеем оценку для k , а ее нельзя получить из соотношения (53.4). Даже если мы имеем обоснованную оценку для k , то оценка (53.5) может быть очень пессимистичной, если окажется, что μ является приближением к хорошо обусловленному собственному значению, а A имеет несколько плохо обусловленных собственных значений.

Однако рассмотрим близкие приближения u_1 и v_1 к правому и левому векторам, соответствующим λ_1 . Для таких u_1 и v_1 можем написать

$$u_1 = k(x_1 + f_1), \quad v_1 = e(y_1 + g_1), \quad (58.1)$$

где f_1 принадлежит подпространству, натянутому на векторы от x_2 до x_n . g_1 — подпространству, натянутому на векторы от y_2 до y_n , и оба они малы. Итак, имеем

$$\begin{aligned} v_1^T A u_1 / v_1^T u_1 &= (y_1^T + g_1^T)(\lambda_1 x_1 + A f_1) / (y_1^T + g_1^T)(x_1 + f_1) = \\ &= (\lambda_1 s_1 + g_1^T A f_1) / (s_1 + g_1^T f_1) = \lambda_1 + \frac{g_1^T (A - \lambda_1 I) f_1}{s_1 + g_1^T f_1}, \end{aligned} \quad (58.2)$$

где

$$s_1 = y_1^T x_1.$$

При этом

$$|v_1^T A u_1 / v_1^T u_1 - \lambda_1| \leq (\|A\|_2 + |\lambda_1|) \|g_1\|_2 \|f_1\|_2 / (|s_1| - \|g_1\|_2 \|f_1\|_2). \quad (58.3)$$

При условии, что s_1 не слишком мало, мы могли бы ожидать, что отношение $(v_1^T A u_1 / v_1^T u_1)$ будет давать хорошее приближение к λ_1 .

Для нормальных матриц отношение Релея $(u_1^H A u_1 / u_1^H u_1)$ имело особую важность. Так как для таких матриц

$$y_1^T = x_1^H, \quad (58.4)$$

то мы видим, что отношение $(v_1^T A u_1 / v_1^T u_1)$ есть естественное расширение отношения Релея для нормальных матриц на ненормальные матрицы. Мы будем называть его *обобщенным отношением Релея, соответствующим v_1 и u_1* .

Предположим теперь, что мы, используя устойчивый алгоритм, нашли нормированные левый и правый собственные векторы v_1 и u_1 , считая их соответствующими одному и тому же собственному значению, и проверили, что для μ , определенного равенством

$$\mu = v_1^T A u_1 / v_1^T u_1, \quad (58.5)$$

как $(Au_1 - \mu_1 u_1)$, так и $(A^T v_1 - \mu_1 v_1)$ малы. Тогда можно с большим основанием предполагать, что $v_1^T u_1$ есть хорошее приближение к s_1 , а μ_1 еще лучшее приближение к λ_1 . Однако без некоторой дальнейшей информации мы не можем получить строгой оценки ошибки в μ_1 или даже быть абсолютно уверенными, что u_1 и v_1 являются приближениями, соответствующими одному и тому же собственному значению.

Анализ ошибок для полной собственной системы

59. Если нам дано приближенное решение полной собственной системы, то при условии, что решение достаточно точное и собственная система не слишком плохо обусловлена, мы можем получить строгие оценки ошибок не только для данной собственной системы, но и для уточненной. Предположим, что μ_i и u_i ($i = 1, \dots, n$) — приближенные собственные значения и собственные векторы матрицы A . Мы можем написать

$$Au_i - \mu_i u_i = r_i, \quad \|u_i\|_2 = 1, \quad (59.1)$$

или, объединяя эти уравнения в матричную форму,

$$AU - U \operatorname{diag}(\mu_i) = R, \quad (59.2)$$

где U и R — матрицы соответственно со столбцами u_i и r_i . Итак,

$$U^{-1}AU = \operatorname{diag}(\mu_i) + U^{-1}R = \operatorname{diag}(\mu_i) + F, \quad (59.3)$$

так что A в точности подобна $\operatorname{diag}(\mu_i) + F$, независимо от того, являются ли μ_i и u_i хорошими приближениями. Однако если μ_i и u_i — хорошие приближения, то матрица R будет иметь малые элементы. Если элементы матрицы F , которая есть решение уравнения

$$UF = R, \quad (59.4)$$

также малы, то правая часть (59.3) будет диагональной матрицей с малым возмущением. В гл. 2, §§ 14–25 мы показали, как свести проблему возмущения для общих матриц к проблеме для диагональных матриц, и большая часть этих параграфов была посвящена проблеме нахождения оценок для собственных значений матрицы вида $\operatorname{diag}(\mu_i) + F$, где F мала. Данные там методы могут быть сразу применены к матрице в правой части (59.3) и, как правило, они могут дать очень точные оценки для всех собственных значений, за исключением патологически близких.

Если нужно полностью реализовать их возможности, то существенно, чтобы R и $U^{-1}R$ были вычислены с некоторым вниманием. Матрица R может быть получена точно на той машине с фиксированной запятой, которая имеет возможность накапливать скалярные произведения, но точное решение (59.4) — трудная задача. Это обсуждается детально в гл. 4, § 69. В общем случае, даже если R задана точно, мы не сможем точно вычислить F .

Однако если мы предположим, что можем вычислить приближенное решение $\bar{F}$ из (59.4) и дать оценку для элементов G , где

$$G = F - \bar{F}, \quad (59.5)$$

то мы имеем

$$U^{-1}AU = \operatorname{diag}(\mu_i) + \bar{F} + G. \quad (59.6)$$

Численный пример

60. В качестве иллюстрации предположим, что для матрицы A порядка 3 мы получили

$$\bar{F} = 10^{-4} \begin{bmatrix} 0,6132 & 0,2157 & -0,4157 \\ 0,2265 & 0,3153 & 0,3123 \\ 0,6157 & 0,2132 & 0,8843 \end{bmatrix}, \quad (60.1)$$

$$\mu = \begin{bmatrix} 0,8132 \\ 0,6132 \\ 0,2157 \end{bmatrix}, \quad |g_{ij}| \leq \frac{1}{2} 10^{-8}. \quad (60.2)$$

Тогда, применяя подобные преобразования с диагональными матрицами, мы можем получить уточненные собственные значения и оценки ошибок для них. Действительно, после умножения первой строки $\text{diag}(\mu_i) + \bar{F} + G$ на 10^{-3} и первого столбца на 10^3 (ср. гл. 2, § 15) круги Гершгорина определяются так:

$$\text{центр } 0,8132 + 10^{-4}(0,6132) + g_{11},$$

$$\text{радиус } 10^{-7}(0,2157 + 0,4157) + 10^{-3}(|g_{12}| + |g_{13}|),$$

$$\text{центр } 0,6132 + 10^{-4}(0,3153) + g_{22},$$

$$\text{радиус } 10^{-1}(0,2265) + 10^{-4}(0,3123) + 10^3|g_{21}| + |g_{23}|,$$

$$\text{центр } 0,2157 + 10^{-4}(0,8843) + g_{33},$$

$$\text{радиус } 10^{-1}(0,6157) + 10^{-4}(0,2132) + 10^3|g_{31}| + |g_{32}|.$$

Вспоминая нашу оценку для g_{ij} , видим, что эти круги разделены и, следовательно, существует собственное значение в круге с центром $0,81326132$ и радиусом $10^{-7}(0,6314) + |g_{11}| + 10^{-3}(|g_{12}| + |g_{13}|)$. Этот радиус не больше чем $10^{-7}(0,6815)$. Аналогичным образом работая с другими строками и столбцами, мы можем улучшить каждое из оставшихся собственных значений и получить оценки ошибок для улучшенных значений. Учитывая анализ главы 2, мы находим, что оценка ошибки для каждого улучшенного μ_i зависит от величины элементов F в i -й строке и i -м столбце и от близости μ_i к другим μ_j .

Условия, ограничивающие достижимую точность

61. Следует подчеркнуть, что $U^{-1}AU$ в точности подобна A , и, вычисляя $\bar{F}$ с достаточной точностью, мы можем быть уверены, что собственные значения матрицы $\text{diag}(\mu_i) + \bar{F}$ будут настолько близки к собственным значениям A , насколько мы захотим. В главе 4 мы опишем метод решения уравнений (59.4) с использованием t -разрядной арифметики с фиксированной запятой, который обычно обеспечивает, чтобы $\bar{F}$ была правильно округленным результатом, так что $|g_{ij}| \leq 2^{-t} \max_{k,l} |f_{kl}|$. Если мы используем такой метод, то оценка для g_{ij} будет непосредственно зависеть от величины F . Нам не нужны для нее априорные оценки, потому что вся информация сама получается после вычислений, однако поучительно рассмотреть, какие факторы определяют величину F .

Если U была матрицей с точными нормированными собственными векторами, то строки U^{-1} будут задаваться отношениями y_i^T/s_i (гл. 2, § 14). Следовательно, мы могли бы ожидать, что в общем случае i -я строка U^{-1} , скажем w_i^T , будет приближением к y_i^T/s_i . Теперь элемент (i, i) матрицы $U^{-1}AU$ есть $w_i^T A u_i$, и так как w_i^T — строка U^{-1} , то имеем

$$w_i^T u_i = 1. \quad (61.1)$$

Поэтому элемент (i, i) может быть представлен в виде $w_i^T A u_i / w_i^T u_i$, и из § 58 следует, что это есть обобщенное отношение Релея, соответствующее w_i и u_i . Это значение уже вычислено и является уточненным собственным значением, полученным по методу §§ 59, 60. Отсюда следует, что основной прием, включенный в этот метод, эквивалентен определению последовательности приближенных левых собственных векторов из данной последовательности приближенных правых собственных векторов, после чего следует вычисление n обобщенных отношений Релея. Организовывая работу так же, как в §§ 59, 60, получим к тому же строгие оценки ошибок.

Так как i -я строка U^{-1} будет приближенно равна y_i^T/s_i , то i -я строка F будет приближенно равна $y_i^T R/s_i$, и мы могли бы ожидать, что эти элементы должны быть большими, если s_i малы. В свою очередь это приводит к более плохим оценкам для g_{ij} . Итак, неудивительно, что малые s_i будут так же ограничивать нашу достижимую точность при заданной точности вычислений, как и близкие собственные значения.

Заметим, что когда мы имеем матрицу U приближенных собственных векторов, то вычисление U^{-1} немедленно дает нам оценки s_i и косвенно, через определение $\|U^{-1}\|_2$, оценку k . Однако теперь мы вынуждены считать оценку для k менее важной, так как метод, который мы дали в § 59, требует меньше работы и дает более полную информацию.

Нелинейные элементарные делители

62. До сих пор мы молчаливо предполагали, что матрица A имеет линейные делители. Когда матрица имеет нелинейные делители, то сразу же возникает дополнительная трудность, если мы используем алгоритм, который основывается на приведении исходной матрицы к одному из более простых видов. Даже если алгоритм вполне устойчив, то вычисленное преобразование будет точно подобно матрице $(A + F)$, где элементы F будут в лучшем случае порядка элементов A , умноженных на 2^{-t} . Как мы видели (гл. 2, § 19), собственные значения, соответствующие нелинейным делителям, могут быть очень чувствительны к возмущениям матрицы и, следовательно, собственные значения преобразованной матрицы могут иметь вполне достаточное разделение и независимые собственные векторы.

Предположим, например, что наша исходная матрица A и ее жорданова каноническая форма задаются так:

$$A = \begin{bmatrix} 1 & 1 & -1 \\ -2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad B' = \begin{bmatrix} 1 & 1 & 0 \\ 10^{-10} & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}. \quad (62.1)$$

Очевидно, что матрица A имеет делители $(\lambda - 3)$ и $(\lambda - 1)^2$. При использовании десятизначного десятичного вычисления, даже очень устойчивый алгоритм может преобразовать A в матрицу, которая точно подобна B' , а не B . Матрица B' имеет простые собственные значения 1 ± 10^{-5} и 3, которые с точки зрения десятизначного вычисления достаточно хорошо разделены, и имеет три независимых собственных вектора. В результате та-

кого вычисления мы могли бы получить собственные значения $1 + 10^{-5}$, $1 - 10^{-5}$, $3 + 10^{-10}$ и соответствующие собственные векторы

$$[u_1 \ u_2 \ u_3] = \begin{bmatrix} 1 & 1 & 10^{-10} \\ 1 + 10^{-5} & 1 - 10^{-5} & 1 + 10^{-10} \\ 1 - 10^{-10} & 1 - 10^{-10} & 1 \end{bmatrix}. \quad (62.2)$$

(Мы предполагаем ошибки порядка 10^{-10} в третьем собственном значении и собственном векторе, так как они хорошо обусловлены.)

Соответствующая матрица невязок такова:

$$[r_1 \ r_2 \ r_3] = \begin{bmatrix} 10^{-10} & 10^{-10} & -10^{-10} - 10^{-20} \\ -2 \cdot 10^{-10} & -2 \cdot 10^{-10} & -4 \cdot 10^{-10} - 10^{-20} \\ -10^{-10} + 10^{-15} & -10^{-10} - 10^{-15} & -2 \cdot 10^{-10} \end{bmatrix}. \quad (62.3)$$

Заметим, что хотя вычисленная собственная система имеет ошибки порядка 10^{-5} , матрица невязок имеет элементы порядка 10^{-10} . Обычно это будет верно для собственных систем, вычисленных по устойчивым методам. Действительно, если λ и u точные для $(A + F)$, вектор-невязка, соответствующий A , есть $-Fu$ и, следовательно, будет зависеть от величины F .

Однако, когда мы пытаемся вычислить $U^{-1}R$, неточность решения обнаруживается сама собой. Действительно,

$$U^{-1}AU = \begin{bmatrix} 1 + 10^{-5} & & \\ & 1 - 10^{-5} & \\ & & 3 + 10^{-10} \end{bmatrix} + F, \quad (62.4)$$

где, обозначая 10^{-5} через x ,

$$F = \begin{bmatrix} -\frac{1}{2}x & -\frac{1}{2}x^2 & -\frac{1}{2}x + x^2 - \frac{1}{2}x^3 & -x - \frac{1}{2}x^2 \\ \frac{1}{2}x + x^2 + \frac{1}{2}x^3 & \frac{1}{2}x & + \frac{1}{2}x^3 & x - \frac{1}{2}x^2 \\ & -2x^2 + x^3 & -2x^2 - x^3 & -x^2 \end{bmatrix} + G; \quad (62.5)$$

$$\left(|g_{ij}| \leq \frac{1}{2} 10^{-15} \right).$$

Мы видим, что некоторые из элементов F будут порядка 10^{-5} . Метод, описанный в гл. 4, § 69, дает решение с оценкой ошибки в этом случае порядка 10^{-15} ; соответственно мы предполагали, что элементы $|G|$ ограничены величиной $\frac{1}{2} 10^{-15}$.

Теорема Гершгорина немедленно дает нам очень точную оценку для третьего собственного значения. При умножении строки 3 на 10^{-5} и столбца 3 на 10^5 круги Гершгорина разделяются и третий имеет

$$\text{центр } (3 + 10^{-10} - 10^{-10} + g_{33}),$$

$$\text{радиус } 10^{-5}(2 \cdot 10^{-10} - 10^{-15} + 2 \cdot 10^{-10} + 10^{-15} + |g_{31}| + |g_{32}|).$$

Итак, мы локализовали собственное значение в круге с центром 3 и радиусом, меньшим чем $6 \cdot 10^{-15}$, несмотря на тот факт, что наша собственная система в значительной степени была неточной. Очевидно, что мы не можем получить столь точных оценок для двух других собственных значений без дальнейших вычислений.

Приближенные инвариантные подпространства

63. Едва ли будет неожиданным, что решение уравнения $UF = R$ много больше, чем само R , так как два из столбцов матрицы U в действительности были приближениями к единственному собственному вектору, соответствующему $\lambda = 1$. Чем больше точность, с которой мы работаем, тем ближе к вырожденной становится U , но мы получаем все лучшие и лучшие приближения к третьему собственному значению.

Если бы мы заранее знали, что матрица имеет квадратичный делитель, то было бы более полезно попытаться вычислить *инвариантное подпространство*, соответствующее двойному собственному значению. Говорят, что линейная оболочка, натянутая на векторы

$$v_1, v_2, \dots, v_r,$$

образует инвариантное подпространство A порядка r , если Av_i ($i = 1, \dots, r$) лежат в этом же подпространстве. Простейший вид инвариантного подпространства есть подпространство, натянутое на единственный собственный вектор, соответствующий простому собственному значению. Очевидно, что подпространство, натянутое на корневые векторы, соответствующие нелинейному элементарному делителю, инвариантно. Хотя собственный вектор, соответствующий $\lambda = 1$, чувствителен к возмущениям, инвариантное подпространство второго порядка не чувствительно. Это подпространство натянуто на e и e_2 , так как

$$Ae = e, \quad Ae_2 = e + e_2, \quad (63.1)$$

Каждый из векторов u_1 и u_2 из § 62 может быть представлен в виде $ae + be_2 + O(10^{-10})$ и, следовательно, в пределах точности лежит в подпространстве. Однако отклонение u_1 и u_2 от линейно зависимых определяется вектором, элементы которого лишь $O(10^{-5})$. Так как в общем случае мы должны считать компоненты этих векторов с ошибкой в десятом знаке, то u_1 и u_2 не определяют в действительности подпространство с рабочей точностью. Отметим, что анализ § 59 легко может быть распространен на случай, когда имеем вычисленные приближенные подпространства порядка больше первого. Например, если для матрицы A четвертого порядка

$$\left. \begin{aligned} Au_1 &= \alpha_{11}u_1 + \alpha_{21}u_2 + r_1, & Au_2 &= \alpha_{12}u_1 + \alpha_{22}u_2 + r_2, \\ Au_3 &= \mu_3u_3 + r_3, & Au_4 &= \mu_4u_4 + r_4, \end{aligned} \right\} \quad (63.2)$$

то

$$A[u_1 \ u_2 \ u_3 \ u_4] = [u_1 \ u_2 \ u_3 \ u_4] \begin{bmatrix} \alpha_{11} & \alpha_{12} & & \\ \alpha_{21} & \alpha_{22} & & \\ & & \mu_3 & \\ & & & \mu_4 \end{bmatrix} + [r_1 \ r_2 \ r_3 \ r_4], \quad (63.3)$$

и в прежних обозначениях

$$U^{-1}AU = D + U^{-1}R. \quad (63.4)$$

Помимо случая нелинейных элементарных делителей, нас часто будет интересовать подпространство второго порядка, натянутое на пару комплексно сопряженных собственных векторов, соответствующих вещественной матрице. Если это векторы $u \pm iv$, где u и v — вещественные, то подпространство, натянутое на u и v , инвариантно. Если $\lambda \pm i\mu$ —

соответствующие собственные значения, то

$$\left. \begin{aligned} A(u + iv) &= (\lambda + i\mu)(u + iv) \\ \text{и далее} \quad Au &= \lambda u - \mu v, \\ Av &= \mu u + \lambda v. \end{aligned} \right\} \quad (63.5)$$

Если мы пытаемся вычислить инвариантное подпространство, чтобы заменить u_1, u_2 из § 62, то, возвращаясь к несложной задаче § 62, можем, например, получить

$$[u_1 \ u_2] = \left[\begin{array}{cc|c} 1 + 10^{-10} & 1 - 10^{-10} & \\ -10^{-10} & 2 + 10^{-10} & \\ 1 - 10^{-10} & 1 - 10^{-10} & \end{array} \right], \quad (63.6)$$

и имеем

$$A[u_1 \ u_2] = [u_1 \ u_2] \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & 1\frac{1}{2} \end{array} \right] + \left[\begin{array}{cc} 0 & 2 \cdot 10^{-10} \\ -4 \cdot 10^{-10} & 2 \cdot 10^{-10} \\ -5 \cdot 10^{-10} & 3 \cdot 10^{-10} \end{array} \right]. \quad (63.7)$$

Итак, невязки имеют тот же порядок величины, что и в (62.3). Если мы закончим построение U добавлением прежнего u_3 , то

$$\begin{aligned} A[u_1 \ u_2 \ u_3] &= [u_1 \ u_2 \ u_3] \left[\begin{array}{cc|c} \frac{1}{2} & \frac{1}{2} & \\ -\frac{1}{2} & 1\frac{1}{2} & \\ \hline & & 3 + 10^{-10} \end{array} \right] + \\ &+ \left[\begin{array}{ccc} 0 & 2 \cdot 10^{-10} & -10^{-10} & -10^{-20} \\ -4 \cdot 10^{-10} & 2 \cdot 10^{-10} & -4 \cdot 10^{-10} & -10^{-20} \\ -5 \cdot 10^{-10} & 3 \cdot 10^{-10} & & -2 \cdot 10^{-10} \end{array} \right]. \end{aligned} \quad (63.8)$$

При этом

$$\begin{aligned} U^{-1}AU &= \left[\begin{array}{cc|c} \frac{1}{2} & \frac{1}{2} & \\ -\frac{1}{2} & 1\frac{1}{2} & \\ \hline & & 3 + 10^{-10} \end{array} \right] + \\ &+ \left[\begin{array}{ccc|c} -\frac{1}{2} 10^{-10} + 10^{-20} & 1\frac{1}{2} \cdot 10^{-10} & \frac{1}{2} 10^{-10} - 2\frac{1}{2} 10^{-20} & \\ \frac{1}{2} 10^{-10} + 5 \cdot 10^{-20} & \frac{1}{2} 10^{-10} - 2 \cdot 10^{-20} & -1\frac{1}{2} 10^{-10} + \frac{1}{2} 10^{-20} & \\ -5 \cdot 10^{-10} - 6 \cdot 10^{-20} & 10^{-10} + 4 \cdot 10^{-20} & -10^{-10} + 10^{-20} & \end{array} \right] + G, \end{aligned} \quad (63.9)$$

где $|g_{ij}| \leq \frac{1}{2} 10^{-20}$.

Теперь мы можем заключить третье собственное значение в круг с радиусом порядка 10^{-20} и находимся в лучших условиях, чтобы получить более точные оценки других собственных значений.

Почти нормальные матрицы

64. В гл. 2, § 26, мы показали, что анализ ошибок, основанный на n числах обусловленности, неудовлетворителен для матриц, которые, например, подобны матрице A , где

$$A = \begin{bmatrix} 1 & 10^{-10} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}. \quad (64.1)$$

Эта матрица имеет нелинейный делитель, но собственные значения сравнительно нечувствительны. Естественно спросить, ожидаем ли мы каких-либо особых трудностей, если попытаемся решить собственную проблему для такой матрицы. В действительности такая матрица не вызывает особых трудностей. Преобразованная матрица будет подобна $(A + F)$ и, вообще говоря, не будет иметь нелинейного делителя, а будет иметь два собственных значения $(1 + a \cdot 10^{-10})$ и $(1 + b \cdot 10^{-10})$ и два вполне независимых соответствующих собственных вектора. Конечно, будет чрезвычайно трудно доказать с помощью вычислений, включающих ошибки округления, что исходная матрица имеет нелинейный делитель. Однако следует понимать, что действительно невозможно установить таким методом, что матрица имеет кратное собственное значение, поскольку делаются ошибки округления, хотя часто, когда кратные собственные значения будут целыми числами или корнями многочленов типа $x^2 + px + q$, где p и q — целые, это, очевидно, может быть сделано численными методами.

Дополнительные замечания

Первый детальный анализ ошибок округления, возникающих при выполнении арифметических операций с фиксированной запятой, был дан Нейманом и Голдстайном (1947) в их фундаментальной работе по обращению матриц. Их система обозначений и общая линия рассмотрения были приняты несколькими последующими авторами.

Соответствующий анализ для операций с плавающей запятой был дан впервые Уилкинсоном в 1957 г. и после периода улучшения опубликован в 1960 г. На выбор формы, в которой выражены результаты, повлияло убеждение, что эффект ошибок округления, сделанных в ходе решения матричных проблем, лучше всего выражается в терминах эквивалентных возмущений исходной матрицы. Результаты § 55 могут быть несколько улучшены следующим образом. Для любого μ имеем

$$Av - \mu v = Av - \mu_R v + (\mu_R - \mu)v = \eta_R + (\mu_R - \mu)v, \quad (1)$$

откуда

$$\|Av - \mu v\|_2^2 = \varepsilon^2 + |\mu_R - \mu|^2 v^2 \quad (2)$$

из ортогональности η_R и v . Следовательно, существует по крайней мере одно собственное значение в круге с центром μ и радиусом y . Рассмотрим теперь значения μ также, что

$$y + |\mu_R - \mu| = a. \quad (3)$$

Для таких μ соответствующие круги и круг с центром μ_R и радиусом a имеют внутреннее касание. Следовательно, единственное собственное значение из последнего круга лежит в каждом из соответствующих кругов с центром μ , удовлетворяющим (3), и поэтому в их пересечении. Очевидно, что пересечение есть круг с центром μ_R и радиусом $y - |\mu_R - \mu|$, и мы имеем

$$y - |\mu_R - \mu| = (y^2 - |\mu_R - \mu|^2) / (y + |\mu_R - \mu|) = \varepsilon^2 / a. \quad (4)$$

Поэтому единственное собственное значение находится в круге с центром μ_R и радиусом ε^2/a . Этот результат неявно присутствует в работе Бауэра и Хаусхолдера (1960). Более точный результат для эрмитовых матриц был дан Като (1949) и Темпле (1952). На практике присутствие дополнительного множителя $(1 - \varepsilon^2/a^2)$ в (55.3) не очень важно, так как результат имеет значение лишь тогда, когда a существенно больше, чем ε .

РЕШЕНИЕ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Введение

1. Детальное сравнение численных методов решения систем линейных уравнений, обращения матриц и вычисления определителей было дано в одной из книг этой серии (Фокс, 1964). Но так как мы время от времени будем заниматься этими задачами, нам нужно их обсудить. Большинство из систем, которые мы должны будем решать, очень плохо обусловлено, и важно, чтобы значение этого факта было полностью оценено. Изучение систем линейных уравнений даст также возможность приобрести некоторый опыт в анализе ошибок и рассмотрении устойчивости на более простых односторонних эквивалентных преобразованиях, прежде чем переходить к подобным преобразованиям. Поэтому сначала рассмотрим те факторы, которые определяют чувствительность решения системы уравнений

$$Ax = b \quad (1.1)$$

к изменению матрицы A и правой части b . Будем иметь дело лишь с такими системами, у которых матрица A квадратная.

Теория возмущений

2. Рассмотрим сначала чувствительность решения (1.1) к возмущению в b . Если b изменяется до $b + k$ и

$$A(x + h) = b + k, \quad (2.1)$$

то, вычитая (1.1), имеем

$$Ah = k, \quad (2.2)$$

так что

$$h = A^{-1}k \quad (2.3)$$

и, следовательно,

$$\frac{\|h\|}{\|k\|} \leq \|A^{-1}\|. \quad (2.4)$$

Обычно мы будем интересоваться относительной ошибкой, и для нее

$$\|h\|/\|x\| \leq \|A^{-1}\| \|k\|/\|A^{-1}b\| = \|A\| \|A^{-1}\| \|k\|/\|b\|, \quad (2.5)$$

так что $\|A\| \|A^{-1}\|$ может рассматриваться как число обусловленности для этой задачи. Если $\|A\| \|A^{-1}\|$ очень велико, то этот результат будет очень пессимистичен для большинства правых частей b и возмущений k . При этом большинство b будут такими, что

$$\|x\| \gg \|A^{-1}\| \|b\|, \quad (2.6)$$

но всегда существуют такие b и k , для которых (2.5) реально.

3. Обращаясь к эффекту возмущения A , можем написать

$$(A + F)(x + h) = b, \quad (3.1)$$

откуда

$$(A + F)h = -Fx. \quad (3.2)$$

Даже если A невырожденная (а мы, конечно, предполагаем это), $(A + F)$ может быть вырожденной, если на F не наложено никаких ограничений. Записывая

$$A + F = A(I + A^{-1}F), \quad (3.3)$$

видим, что $(A + F)$ будет невырожденной, если

$$\|A^{-1}F\| < 1. \quad (3.4)$$

Предполагая выполненным это условие, имеем

$$h = -(I + A^{-1}F)^{-1}A^{-1}Fx \quad (3.5)$$

и, следовательно,

$$\|h\| \leq \frac{\|A^{-1}F\| \|x\|}{1 - \|A^{-1}F\|} \leq \frac{\|A^{-1}\| \|F\| \|x\|}{1 - \|A^{-1}\| \|F\|}, \quad (3.6)$$

если только

$$\|A^{-1}\| \|F\| < 1. \quad (3.7)$$

Для вычисления относительной ошибки нам интересна какая-нибудь оценка для $\|h\|/\|x\|$ через $\|F\|/\|A\|$. Мы можем записать (3.6) в виде

$$\|h\|/\|x\| \leq \|A\| \|A^{-1}\| \frac{\|F\|}{\|A\|} / \left(1 - \|A\| \|A^{-1}\| \frac{\|F\|}{\|A\|}\right), \quad (3.8)$$

и опять величина $\|A\| \|A^{-1}\|$ является решающей. В качестве числа обусловленности наиболее часто используется $\|A\|_2 \|A^{-1}\|_2$. Обычно оно обозначается через $k(A)$ и называется *спектральным числом обусловленности A для задачи обращения*. Теперь мы видим из гл. 2, §§ 30, 31, что спектральное число обусловленности матрицы A для ее проблемы собственных значений есть спектральное число обусловленности ее матрицы собственных векторов для задачи обращения.

При выводе соотношений (3.6) и (3.8) мы заменили $\|A^{-1}F\|$ на $\|A^{-1}\| \|F\|$. Всегда будут существовать такие возмущения, при которых эта замена пессимистична. Если, например, $F = \alpha A$, то для (3.6)

$$\|h\| \leq \frac{|\alpha| \|x\|}{1 - |\alpha|}, \quad (3.9)$$

и видно, что число обусловленности сюда не входит.

Числа обусловленности

4. Спектральное число обусловленности k инвариантно в отношении эквивалентных унитарных преобразований. Если R унитарная и

$$B = RA, \quad (4.1)$$

то мы имеем

$$\|B\|_2 \|B^{-1}\|_2 = \|RA\|_2 \|A^{-1}R^H\|_2 = \|A\|_2 \|A^{-1}\|_2. \quad (4.2)$$

Аналогично, если c — любое число, то, очевидно,

$$k(cA) = k(A). \quad (4.3)$$

Это разумный результат, так как мы не могли ожидать изменения обусловленности системы уравнений от умножения обеих частей на константу. Из определения числа обусловленности следует, что

$$k(A) = \|A\|_2 \|A^{-1}\|_2 \geq \|AA^{-1}\|_2 = 1 \quad (4.4)$$

для любой матрицы, причем равенство достигается, если A кратна унитарной матрице.

Если σ_1 и σ_n являются наибольшим и наименьшим главными значениями A (гл. 1, § 53), то

$$k(A) = \frac{\sigma_1}{\sigma_n}, \quad (4.5)$$

тогда как, если A — симметричная, то

$$k(A) = \frac{|\lambda_1|}{|\lambda_n|}, \quad (4.6)$$

где λ_1 и λ_n являются наибольшим и наименьшим по модулю собственными значениями. Появление отношений в равенствах (4.5) и (4.6) часто приводит в недоумение, но важно понять, что числители — всего лишь нормирующие множители.

Если мы возьмем $\|A\|_2 = 1$, что всегда может быть достигнуто умножением системы уравнений $Ax = b$ на подходящую константу, то

$$k(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{1}{\sigma_n}, \quad (4.7)$$

и если A симметричная, то

$$k(A) = \frac{1}{|\lambda_n|}. \quad (4.8)$$

Поэтому нормированная матрица A плохо обусловлена лишь тогда, когда $\|A^{-1}\|_2$ велика, т. е. когда σ_n мало или, в случае симметричной матрицы, когда λ_n мало. При использовании k в качестве числа обусловленности заключаем, что нормированная матрица плохо обусловлена, если обратная имеет большие элементы, что согласуется с принятым понятием плохой обусловленности, так как это означает, что малое изменение правой части может привести к большому изменению в решении.

Уравновешенные матрицы

5. Ортогональная матрица не только нормирована, но и все ее строки и столбцы единичной длины. Очевидно, что это неверно для нормированных матриц вообще. Предположим, что мы умножаем r -й столбец ортогональной матрицы A на 10^{-10} и получаем матрицу B . Матрица $B^T B$ будет единичной матрицей, за исключением ее r -го диагонального элемента, который равен 10^{-20} , следовательно, $\|B\|_2 = 1$. Обратная матрица от B есть A^T с ее r -й строкой, умноженной на 10^{10} . Следовательно, $B^{-1} (B^{-1})^T$ есть диагональная матрица с единичными элементами, за исключением

r -го диагонального элемента, который равен 10^{20} . Поэтому $\|B^{-1}\|_2 = 10^{10}$ и

$$k(B) = \|B\|_2 \|B^{-1}\|_2 = 10^{40}. \quad (5.1)$$

Мы получим тот же результат, если умножим любую строку ортогональной матрицы на 10^{-10} .

Матрица, у которой каждая строка и столбец имеют длину порядка единицы, будет называться *уравновешенной* матрицей. Мы не будем использовать этот термин в очень точном смысле, тем не менее наиболее естественно рассмотреть матрицу, у которой каждый элемент меньше единицы и все строки и столбцы нормированы так, что по крайней мере модуль одного элемента находится между $1/2$ и 1 . Чтобы избежать ошибок округления при нормировании, обычно используют степени 2 или 10. Если матрица симметричная, то выгодно сохранить симметрию при нормировании и, следовательно, нужно удовлетворить требованию, чтобы каждый столбец содержал элемент, по модулю больший чем $1/4$. Если любая строка или столбец уравновешенной матрицы A умножается на 10^{-10} , то полученная матрица B должна иметь очень большое число обусловленности. Наибольшее собственное значение σ_1^2 матрицы BB^T больше наибольшего ее диагонального элемента, а ее наименьшее собственное значение σ_n^2 меньше наименьшего диагонального элемента. Следовательно, σ_1^2 больше чем $1/4$, и σ_n^2 меньше чем $10^{-20} n$, так что

$$k(B) \geq \frac{10^{40}}{(2n)^{1/2}}. \quad (5.2)$$

Заметим, что этот результат верен независимо от того, каким может быть число обусловленности A .

6. Естественно спросить, действительно ли матрица B , полученная из хорошо обусловленной матрицы A , «плохо обусловлена». Трудность ответа на этот вопрос заключается в том, что термин «плохо обусловленная» используется как в общепринятом, так и в точно определенном математическом смысле. Если мы *определяем* k как число обусловленности матрицы, то без сомнения измененная ортогональная матрица из § 5 очень «плохо обусловлена». Однако в общепринятом смысле мы говорим, что система уравнений плохо обусловлена, если вычисленное решение имеет небольшую практическую ценность. Мы не можем решить, плохо ли обусловлены уравнения, без исследования пути получения коэффициентов.

Простые практические примеры

7. Рассмотрим простой пример, в котором коэффициенты системы уравнений второго порядка находятся следующим образом. Каждый из четырех элементов матрицы A определяется измерением трех независимых величин, скажем, каждая из которых лежит в пределах ± 1 , и последующим их сложением между собой. Предположим, что ошибки индивидуальных измерений порядка 10^{-5} и что полученная матрица такова:

$$\begin{bmatrix} 1,00000 & 0,00001 \\ 1,00000 & -0,00001 \end{bmatrix}, \quad (7.1)$$

причем при вычислении элементов второго столбца имело место сокращение. Если мы умножим второй столбец на 10^5 , то эта матрица с точностью

до множителя $2^{1/2}$ будет ортогональной. Система уравнений, очевидно, плохо обусловлена, в том смысле, что вычисленным ответам нельзя верить.

Рассмотрим теперь вторую ситуацию, в которой элементы матрицы второго порядка получаются из измерения одной величины с относительной точностью 10^{-5} . Предположим, что элементы второго столбца измерены в единицах, которые много больше, чем единицы, использованные в первом столбце. Мы могли бы опять получить матрицу (7.1), но теперь вычисленные решения могут быть вполне удовлетворительны.

Аналогичные рассуждения относятся и к тому случаю, когда элементы матрицы вычисляются из математических выражений по методам, которые содержат ошибки округления. Вообще говоря, будет существенная разница между ситуациями, в которых матрица типа (7.1) имела малые элементы во втором столбце в результате сокращения близких по модулю и противоположных по знаку чисел, и ситуациями, в которых эти элементы имеют большую относительную точность.

Обусловленность матрицы собственных векторов

8. В главе 2 мы интересовались обусловленностью матрицы со столбцами, равными нормированным правосторонним собственным векторам заданной матрицы. Такая матрица не может иметь *столбец* из малых элементов, но возможно, что вся *строка* будет малой. Если это так, то собственные векторы, конечно, будут почти линейно зависимы, а соответствующая проблема собственных значений плохо обусловлена. Очевидно, будет «правильно» считать такую матрицу собственных векторов плохо обусловленной. Даже в этом случае есть много факторов за уравнивание матрицы умножением строки с малыми элементами на степень 2 (или 10), прежде чем обратить ее для того, чтобы найти левостороннюю систему собственных векторов. После этого для получения верной обратной матрицы следует в вычисленной обратной умножить соответствующий столбец на ту же степень 2 (или 10).

9. Мы видим теперь, что обусловленность системы линейных алгебраических уравнений в общепринятом смысле есть вещь достаточно субъективная. Но что важно на практике, так это обусловленность численной задачи, связанной с решением системы.

Некоторые методы, обсуждаемые в этой главе, более пригодны для уравновешенных матриц, потому что время от времени принимаются решения, зависящие от относительной величины элементов матрицы. Так как мы почти всегда уменьшаем число обусловленности матрицы уравниванием, то доводы в его пользу кажутся очень убедительными. Если мы используем арифметику с плавающей запятой, то нормировка, в отличие от уравнивания, не играет какой-либо существенной роли. Поэтому предположение о нормировке упрощает исследование без особой потери общности.

Если матрица уравновешена и нормирована в строгом смысле, то даже если мы рассматриваем и симметричное уравнивание, $\|A\|_2$ лежит между $1/4$ и n и будет обычно порядка $n^{1/2}$. Итак, за исключением очень больших матриц, $\|A^{-1}\|_2$ (или даже максимальный по модулю элемент A^{-1}) будет приемлемым числом обусловленности. Так как r -й столбец A^{-1} есть решение системы $Ax = e_r$, то очевидно, что нормированная уравновешенная матрица плохо обусловлена тогда и только тогда, когда существует такой нормированный вектор b , для которого решение системы $Ax = b$ имеет большую норму.

Явное решение

10. Если A имеет линейные элементарные делители, то можем получить простое явное решение уравнения $Ax = b$. Если мы напишем

$$x = \sum \alpha_i x_i, \quad (10.1)$$

где x_i образуют полную систему нормированных собственных векторов A , то

$$\sum \lambda_i \alpha_i x_i = b. \quad (10.2)$$

Следовательно, если y_i являются соответствующими собственными векторами A^T , то

$$\alpha_i = y_i^T b (\lambda_i y_i^T x_i)^{-1} = y_i^T b (\lambda_i s_i)^{-1}. \quad (10.3)$$

Из (10.1) и (10.3) имеем

$$\|x\| \leq \sum |\alpha_i| \|x_i\| = \sum |\alpha_i| \leq \|b\| \sum |\lambda_i s_i|^{-1}. \quad (10.4)$$

Мы не можем получить большое решение, если по крайней мере одно из λ_i или s_i (или оба) не мало. Если A — симметричная, то все s_i равны единице, так что A не может быть плохо обусловленной до тех пор, пока, согласно полученному выше результату, у нее не будет малого собственного значения.

Общие замечания об обусловленности матриц

11. Важно заметить, что нормированная несимметричная матрица может быть очень плохо обусловленной, не имея малых собственных значений. Например, система уравнений порядка 100

$$\begin{bmatrix} 0,501 & -1 & & & \\ & 0,502 & -1 & & \\ & & \dots & \dots & \\ & & & 0,599 & -1 \\ & & & & 0,600 \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (11.1)$$

для первой компоненты своего решения имеет

$$x_1 = 1/(0,600 \times 0,599 \times \dots \times 0,501) > (0,6)^{-100} > 10^{22}. \quad (11.2)$$

Следовательно, норма матрицы, обратной к матрице коэффициентов, будет, конечно, больше чем 10^{22} . Несмотря на это, ее наименьшее собственное значение есть 0,501. Эта матрица должна иметь главное значение порядка 10^{-22} .

С другой стороны, наличие малого собственного значения нормированной матрицы обязательно означает плохую обусловленность, так как

$$\|A^{-1}\|_2 \geq |\max \text{собств. знач. } A^{-1}| = 1/\min |\lambda_i|. \quad (11.3)$$

Однако наличие малого s_i не обязательно означает плохую обусловленность. Это иллюстрируется матрицей (26.6) из главы 2. Она имеет s_i порядка 10^{-10} , но ее обратная матрица такова:

$$\begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & -1/(1+x) \\ 0 & 0 & 1/(1+x) \end{bmatrix} \quad (x = 10^{-10}). \quad (11.4)$$

Обращаясь к матрицам с нелинейными делителями, мы видим на примере матрицы

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad (11.5)$$

что их присутствие не обязательно означает плохую обусловленность. С другой стороны, если мы заменим все диагональные элементы матрицы (11.1) на 0,500, то она станет еще более плохо обусловленной и будет иметь элементарный делитель степени 100.

Эти примеры показывают, что хотя обусловленности матрицы в отношении проблемы собственных значений и в отношении обращения имеют совсем различное содержание, все же они в некотором смысле связаны.

Связь плохой обусловленности и почти вырожденности

12. Если матрица вырождается, то она должна иметь по крайней мере одно нулевое собственное значение, так как произведение ее собственных значений равно определителю. Естественно представить себе вырожденность как предельную форму плохой обусловленности, и поэтому заманчиво сделать вывод, что плохо обусловленная нормированная матрица должна иметь малое собственное значение. Мы только что видели, что это неверно, однако поучительно посмотреть, почему представление через предельный процесс так обманчиво.

Теперь покажем, что если n фиксировано и $\|A\|_2 = 1$, то наименьшее по модулю собственное значение λ_n обязательно стремится к нулю, когда $\|A^{-1}\|_2$ стремится к бесконечности, но при этом мы сможем гарантировать лишь очень медленную скорость сходимости. Для любой матрицы A существует унитарная R такая, что

$$R^H A R = \text{diag}(\lambda_i) + T = D + T, \quad (12.1)$$

где T — строго верхняя треугольная (ср. гл. 3, § 47). Следовательно,

$$R^H A^{-1} R = (D + T)^{-1} = [I - K + K^2 - \dots (-1)^{n-1} K^{n-1}] D^{-1}, \quad (12.2)$$

где $K = D^{-1} T$ и $K^r = 0$ ($r \geq n$), потому что K — строго верхняя треугольная матрица. Для нашей цели будет пригодна совсем грубая оценка. Мы имеем

$$\|T\|_2 \leq \|T\|_E \leq \|D + T\|_E = \|A\|_E \leq n^{1/2} \|A\|_2 = n^{1/2} \quad (12.3)$$

и

$$\|K\|_2 \leq \|D^{-1}\|_2 \|T\|_2 \leq n^{1/2} / |\lambda|, \quad (12.4)$$

где λ есть наименьшее по модулю собственное значение A . Итак, из (12.2)

$$\begin{aligned} A^{-1} \|_2 &= \|R^H A^{-1} R\|_2 \leq \\ &\leq [1 + n^{1/2}/|\lambda| + (n^{1/2}/|\lambda|)^2 + \dots + (n^{1/2}/|\lambda|)^{n-1}] / |\lambda| \leq \\ &\leq n^{\frac{1}{2}(n+1)} / |\lambda|^n, \end{aligned} \quad (12.5)$$

откуда

$$|\lambda| \leq (n^{1/2(n+1)} / \|A^{-1}\|_2)^{1/n}. \quad (12.6)$$

Это, конечно, показывает, что при фиксированном n собственное значение λ стремится к нулю, когда $\|A^{-1}\|_2$ стремится к бесконечности, но как

корень n -й степени. Примеры, которые, мы обсудили, показывают, что в отношении знаменателя $\|A^{-1}\|_2^{1/n}$ оценка вполне реальна, и мы можем построить нормированную матрицу порядка 100, имеющую число обусловленности около 2^{100} , но с собственными значениями, большими чем 0,5.

Ограничения, накладываемые t -разрядной арифметикой

13. Если A и b системы

$$Ax = b \quad (13.1)$$

требуют для своего представления более чем t разрядов, то мы должны довольствоваться t -разрядной аппроксимацией A' и b' . Имеем

$$A' = A + F, \quad b' = b + k, \quad (13.2)$$

где для вычислений с фиксированной запятой

$$|f_{ij}| \leq \frac{1}{2} 2^{-t}, \quad |k_i| \leq \frac{1}{2} 2^{-t}, \quad (13.3)$$

так что

$$\|F\|_2 \leq \frac{1}{2} n 2^{-t}, \quad \|k\|_2 \leq \frac{1}{2} n^{1/2} 2^{-t}. \quad (13.4)$$

Очевидно, что верхние оценки могут быть достигнуты. Если только $\frac{1}{2} n \|A^{-1}\|_2 2^{-t} < 1$, то, согласно (3.7), матрица $(A + F)$ не может быть вырожденной. Если это условие выполнено, то простые вычисления показывают, что точное решение x' системы $A'x' = b'$ удовлетворяет соотношению

$$\|x' - x\|_2 \leq (\|A^{-1}\|_2 \|F\|_2 \|x\|_2 + \|A^{-1}\|_2 \|k\|_2) / (1 - \|A^{-1}\|_2 \|F\|_2). \quad (13.5)$$

Очевидно, что если $n 2^{-t} \|A^{-1}\|_2$ незначительно меньше единицы, то начальные ошибки могут полностью исказить точное решение. Для вычислений с плавающей запятой аналогично имеем

$$\left. \begin{aligned} |f_{ij}| &\leq 2^{-t} |a_{ij}|, \quad |k_i| \leq 2^{-t} |b_i|, \\ \|F\|_2 &\leq \|F\|_E \leq 2^{-t} \|A\|_E \leq 2^{-t} n^{1/2} \|A\|_2, \end{aligned} \right\} \quad (13.6)$$

$$\|k\|_2 \leq 2^{-t} \|b\|_2. \quad (13.7)$$

14. Обратимся теперь к ошибкам округления, сделанным во время вычислений. Рассмотрим результат умножения системы уравнений на константу c , лежащую между $1/2$ и 1 и требующую t разрядов для своего представления в арифметике с фиксированной запятой. Если мы округлим полученные числа, то преобразованная система уравнений такова:

$$\left. \begin{aligned} (cA + F)x &= cb + k, \\ |f_{ij}| &\leq \frac{1}{2} 2^{-t}, \quad |k_i| \leq \frac{1}{2} 2^{-t}. \end{aligned} \right\} \quad (14.1)$$

Эта система в точности эквивалентна системе

$$(A + F/c)x = b + k/c, \quad (14.2)$$

так что ошибки округления эквивалентны возмущениям F/c и k/c соответственно в A и b , причем их оценки в $1/c$ раз больше оценок первоначаль-

ных ошибок. Типичный шаг большинства прямых методов состоит в умножении одного из уравнений на число и прибавлении к другому уравнению. Так как даже для матриц умеренного порядка выполняется много таких шагов, мы могли бы ожидать, что вся совокупность ошибок округления ухудшит точность решения значительно более серьезно, чем единственная операция (14.1). Как это ни удивительно, один из самых простых методов решения обычно приводит к ошибке, предполагаемое значение которой точно такое, какое получается при случайных возмущениях F и k , удовлетворяющих соотношениям (13.3). Это означает, что если исходные числа не являются точными числами из t двоичных разрядов, то ошибки, получающиеся от любого начального округления, могут быть настолько же серьезны, как и ошибки, возникающие при решении на всех шагах. Более неожиданным является тот факт, что для некоторых очень плохо обусловленных матриц начальное округление имеет даже значительно более серьезное значение (см. §§ 40 и 45).

Алгоритмы решения линейных уравнений

15. Будем интересоваться лишь прямыми методами решения линейных уравнений. Все алгоритмы, которые мы используем, основаны на следующем преобразовании. Если x — решение системы

$$Ax = b, \quad (15.1)$$

где A есть $(n \times n)$ -матрица, то оно будет также решением системы

$$PAx = Pb, \quad (15.2)$$

где P — любая $(m \times n)$ -матрица. Однако если P — квадратная и невырожденная, то любое решение (15.2) является решением (15.1), и наоборот. Далее, если Q — также невырожденная и y есть решение

$$PAQy = Pb, \quad (15.3)$$

то Qy есть решение (15.1) (ср. гл. 1, § 15). Некоторые из наших алгоритмов будут основаны на использовании этого последнего результата, причем в качестве Q берется матрица перестановок (гл. 1, § 40 (ii)).

Основной прием состоит в определении такой простой невырожденной матрицы P , чтобы PA была *верхней треугольной*. Тогда система уравнений (15.2) может быть решена сразу же, так как n -е уравнение содержит лишь x_n , $(n-1)$ -е содержит x_n и x_{n-1} и т. д. Поэтому неизвестные могут быть определены в порядке $x_n, x_{n-1}, \dots, x_1$. Метод решения системы уравнений с верхней треугольной матрицей коэффициентов обычно называется *обратной подстановкой или обратным ходом*.

Приведение A к треугольному виду имеет большое значение и в том случае, когда не требуется обратная подстановка, но нужна триангуляризация.

Алгоритмы, которые мы рассмотрим, обычно состоят из $(n-1)$ основных шагов и имеют следующие общие особенности. Обозначим первоначальную систему уравнений через

$$A_0x = b_0. \quad (15.4)$$

Каждый шаг приводит к получению новой системы уравнений, которая эквивалентна исходной. Обозначим p -ю эквивалентную систему через

$$A_px = b_p \quad (15.5)$$

Существенная особенность алгоритмов заключается в том, что первые p столбцов A_p будут столбцами треугольной матрицы; например, для $n = 6$, $p = 3$ уравнения имеют вид

$$\begin{matrix} p \\ n-p \end{matrix} \left\{ \begin{matrix} \overbrace{\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}}^p & \overbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}^{n-p} \end{matrix} \right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \times \\ \times \\ \times \\ \times \\ \times \\ \times \end{bmatrix}. \quad (15.6)$$

Здесь крестики означают элементы, которые, вообще говоря, ненулевые. Обычно опускают столбец x_i и знак равенства и оставляют лишь матрицу $(A_p \parallel b_p)$.

Теперь r -й основной шаг состоит в определении элементарной матрицы P_r такой, что A_r , определяемая матрицей $P_r A_{r-1}$, будет верхней треугольной в своих первых r столбцах и будет иметь те же первые $(r-1)$ строк, что и матрица A_{r-1} . Поэтому этот шаг оставляет первые $(r-1)$ уравнений без изменения. Уравнения, определяющие r -е преобразование, таковы:

$$A_r = P_r A_{r-1}, \quad b_r = P_r b_{r-1}. \quad (15.7)$$

Очевидно, что r -й столбец матрицы $P_r A_{r-1}$ равен

$$P_r \begin{bmatrix} a_{1,r}^{(r-1)} \\ a_{2,r}^{(r-1)} \\ \vdots \\ a_{n,r}^{(r-1)} \end{bmatrix}, \quad (15.8)$$

и P_r должна быть выбрана так, чтобы элементы от $a_{1,r}^{(r-1)}$ до $a_{r-1,r}^{(r-1)}$ не изменялись, а $a_{r+1,r}^{(r-1)}, \dots, a_{n,r}^{(r-1)}$ стали нулями. Это та же самая задача, которую мы рассматривали в связи с нашим основным анализом ошибок в главе 3. Большая часть из оставшегося в этой главе посвящена обсуждению соответствующих алгоритмов.

Метод Гаусса

16. Самый простой алгоритм получается, если мы возьмем в качестве матрицы P_r элементарную матрицу N_r (гл. 1, § 40) с элементами

$$n_{ir} = a_{ir}^{(r-1)} / a_{rr}^{(r-1)}. \quad (16.1)$$

Знаменатель $a_{rr}^{(r-1)}$ называется *опорным или ведущим элементом r -го шага*. Левое умножение на N_r приводит к вычитанию r -й строки, умноженной на n_{ir} , из i -й строки для всех значений i от $(r+1)$ до n , причем множители выбраны так, чтобы исключить последние $(n-r)$ элементов в r -м столбце. Те же самые операции выполняются над правой частью b_{r-1} . Очевидно, что строки от 1-й до $(r-1)$ -й не изменяются и, кроме того, не изменяется ни один из нулей в первых $(r-1)$ столбцах, так как каждый из них заменяется линейной комбинацией нулевых элементов. По этому алгоритму не изменяется также и r -я строка. Читатель хорошо

знает этот алгоритм как обычный вариант метода исключения, изучаемого в школьной алгебре. Действительно, r -е уравнение используется для исключения x_r из уравнений от $(r+1)$ до n системы $A_{r-1}x = b_{r-1}$. Этот алгоритм обычно называется *методом Гаусса*.

Треугольное разложение

17. Объединяя уравнения (15.7) с $P_r = N_r$ для r от 1 до $(n-1)$, имеем

$$N_{n-1} \dots N_2 N_1 A_0 = A_{n-1}, \quad N_{n-1} \dots N_2 N_1 b_0 = b_{n-1}. \quad (17.1)$$

Первое из уравнений дает

$$A_0 = N_1^{-1} N_2^{-1} \dots N_{n-1}^{-1} A_{n-1} = N A_{n-1}, \quad (17.2)$$

где N есть нижняя треугольная матрица с единичными диагональными элементами и поддиагональными элементами, равными n_{ij} и определенными в (16.1). В произведение N_i^{-1} не входит ни одно произведение n_{ij} . Итак, метод Гаусса дает нижнюю треугольную матрицу N с единичными диагональными элементами и верхнюю треугольную матрицу A_{n-1} такие, что их произведение равно A_0 .

Если A_0 может быть представлена таким произведением (ср. § 20) и невырождена, то представление единственно. Пусть

$$A_0 = L_1 U_1 = L_2 U_2, \quad (17.3)$$

где L_i — нижние треугольные матрицы с единичными диагональными элементами и U_i — верхние треугольные; так как

$$\det(A_0) = \det(L_i) \det(U_i), \quad (17.4)$$

U_i не могут быть вырожденными. Следовательно, (17.3)₁ дает

$$L_2^{-1} L_1 = U_2 U_1^{-1}. \quad (17.5)$$

Матрица в левой части есть произведение двух нижних треугольных матриц с единичными диагональными элементами и поэтому сама такая же, тогда как матрица в правой части верхняя треугольная. Итак, оба произведения должны быть единичными матрицами, и $L_1 = L_2$, $U_1 = U_2$.

Структура матриц треугольного разложения

18. Для дальнейших приложений имеют большое значение соотношения между различными подматрицами A_0 , N , A_{n-1} и промежуточными A_r . Обозначим соответствующие разбиения этих матриц через

$$A_0 = \left[\begin{array}{c|c} A_{rr} & A_{r, n-r} \\ \hline A_{n-r, r} & A_{n-r, n-r} \end{array} \right], \quad N = \left[\begin{array}{c|c} L_{rr} & 0 \\ \hline L_{n-r, r} & L_{n-r, n-r} \end{array} \right], \quad (18.1)$$

$$A_{n-1} = \left[\begin{array}{c|c} U_{rr} & U_{r, n-r} \\ \hline 0 & U_{n-r, n-r} \end{array} \right], \quad A_r = \left[\begin{array}{c|c} U_{rr} & U_{r, n-r} \\ \hline 0 & W_{n-r, n-r} \end{array} \right]. \quad (18.2)$$

Здесь матрица с индексами i, j имеет i строк и j столбцов. Мы будем обозначать N и A_{n-1} соответственно через L и U , если не хотим подчеркнуть их связь с выполнением алгоритма. Обозначения в (18.2) отражают тот факт, что первые r строк A_r не изменяются на последующих шагах; однако

мы видели, что $(r + 1)$ -я строка A_r также не изменяется на этих шагах, так что первые строки $U_{n-r, n-r}$ и $W_{n-r, n-r}$ совпадают.

Из первых r основных шагов имеем

$$N_r N_{r-1} \dots N_2 N_1 A_0 = A_r,$$

откуда

$$A_0 = N_1^{-1} N_2^{-1} \dots N_r^{-1} A_r = \left[\begin{array}{c|c} L_{rr} & 0 \\ \hline L_{n-r, r} & I \end{array} \right] A_r, \quad (18.3)$$

учитывая свойство N_i в отношении перемножения (гл. 1, § 41). Приравнивая клетки в (18.2), получаем

$$\left. \begin{aligned} A_{rr} &= L_{rr} U_{rr}, & A_{r, n-r} &= L_{rr} U_{r, n-r}, \\ A_{n-r, r} &= L_{n-r, r} U_{rr}, & A_{n-r, n-r} &= \\ & & &= L_{n-r, r} U_{r, n-r} + L_{n-r, n-r} U_{n-r, n-r}, \end{aligned} \right\} \quad (18.4)$$

тогда как из последнего разбиения в (18.3) находим

$$A_{n-r, n-r} = L_{n-r, r} U_{r, n-r} + W_{n-r, n-r}. \quad (18.5)$$

Эти последние пять уравнений дают нам значительное количество информации. Первое показывает, что $L_{rr} U_{rr}$ есть треугольное разложение A_{rr} , главной подматрицы порядка r матрицы A_0 . Поэтому, осуществляя метод Гаусса, получаем по очереди треугольное разложение каждой из главных подматриц A_0 . Далее, треугольные матрицы, соответствующие r -й главной подматрице A_0 , являются главными подматрицами окончательных треугольных матриц N и A_{n-1} . Последнее уравнение из (18.4) и (18.5) показывает, что

$$L_{n-r, n-r} U_{n-r, n-r} = W_{n-r, n-r}, \quad (18.6)$$

так что $L_{n-r, n-r} U_{n-r, n-r}$ есть треугольное разложение квадратной матрицы порядка $(n - r)$ в нижнем правом углу матрицы A_r .

Явные выражения для элементов треугольных матриц

19. Уравнения (18.4) дают возможность получить явные выражения для элементов N и A_{n-1} через миноры A_0 , и они нам потребуются в дальнейшем. Если мы объединим первое и третье уравнения, то получим

$$\begin{bmatrix} A_{rr} \\ A_{n-r, r} \end{bmatrix} = \begin{bmatrix} L_{rr} \\ L_{n-r, r} \end{bmatrix} U_{rr}. \quad (19.1)$$

Приравнивая строки от 1-й до $(r - 1)$ -й и строку i с обеих сторон этого уравнения, имеем

$$A_{ir} = L_{ir} U_{rr} \quad (i = 1, \dots, n), \quad (19.2)$$

где A_{ir} означает матрицу, полученную из первых $(r - 1)$ строк и i -й строки первых r столбцов A_0 ; матрица L_{ir} определяется аналогично. Так как L_{ir} совпадает с L_{rr} в первых $(r - 1)$ строках, то она обязательно треугольная, так что

$$\det(L_{ir}) = l_{ir}. \quad (19.3)$$

Следовательно,

$$\det(A_{ir}) = l_{ir} \det(U_{rr}). \quad (19.4)$$

Для $i = r$

$$\det(A_{rr}) = \det(U_{rr}), \quad (19.5)$$

поэтому

$$n_{ir} = l_{ir} = \det(A_{ir})/\det(A_{rr}) \quad (i = 1, \dots, n). \quad (19.6)$$

Заметим, что $\det(A_{ir}) = 0$ для $i < r$ (так как в этом случае A_{ir} имеет две одинаковые строки) согласно треугольной структуре N .

Аналогично

$$[A_{rr} : A_{r, n-r}] = L_{rr} [U_{rr} : U_{r, n-r}], \quad (19.7)$$

и в соответствующей записи

$$A_{ri} = L_{rr} U_{ri}, \quad (19.8)$$

откуда

$$\det(A_{ri}) = \det(U_{ri}), \quad (i = 1, \dots, n). \quad (19.9)$$

Здесь U_{ri} совпадает с U_{rr} в первых $(r - 1)$ столбцах, и поэтому она обязательно треугольная. Следовательно,

$$\det(A_{ri}) = \det(U_{r-1, r-1}) u_{ri}. \quad (19.10)$$

Объединяя (19.10) и (19.5) с $r - 1$ вместо r , получаем

$$u_{ri} = \det(A_{ri})/\det(A_{r-1, r-1}) \quad (i = 1, \dots, n). \quad (19.11)$$

В частности,

$$\left. \begin{aligned} u_{rr} &= \det(A_{rr})/\det(A_{r-1, r-1}) \quad (r > 1), \\ u_{11} &= a_{11}. \end{aligned} \right\} \quad (19.12)$$

Мы видели раньше в (18.3), что $u_{r+1, r+1}$ есть элемент $(1, 1)$ матрицы $W_{n-r, n-r}$, и уравнение (19.12) показывает, что он равен

$$\det(A_{r+1, r+1})/\det(A_{rr}).$$

Мы можем получить аналогичные выражения для других элементов квадратной матрицы $W_{n-r, n-r}$. Для краткости обозначим элемент (i, j) матрицы A_r через w_{ij} , если $i, j > r$, так что матрица $W_{n-r, n-r}$ составляется из таких элементов. При получении A_r из A_0 лишь прибавляются строки от 1-й до r -й, умноженные на некоторые числа, к другим строкам. Следовательно, любой минор, составленный из первых r строк, не изменяется. Рассмотрим теперь минор $A_{rr, i, j}$, составленный из первых r строк и i -й строки и первых r столбцов и j -го столбца. Этот минор не изменяет своего значения, и из структуры A_r следует, что он равен $\det(U_{rr})w_{ij}$. Следовательно, имеем

$$\det(A_{rr, i, j}) = \det(U_{rr}) w_{ij} \quad (19.13)$$

и из (19.5)

$$w_{ij} = \det(A_{rr, i, j})/\det(A_{rr}). \quad (19.14)$$

Значение, которое мы получили для $u_{r+1, r+1}$, есть просто частный случай (19.14); это не удивительно, так как до того шага, когда $u_{r+1, r+1}$ используется как ведущий элемент, все элементы $W_{n-r, n-r}$ преобразуются сходным образом. Теперь имеем явное выражение для всех элементов N и A_r .

Несостоятельность метода Гаусса

20. Выражения, которые мы получили для элементов N и A_r , основываются на предположении, что метод Гаусса в действительности может быть осуществлен. Однако алгоритм не может не осуществиться, если все ведущие элементы ненулевые, а из (19.13) следует, что это верно, если все главные миноры от 1-го до $(n-1)$ -го порядка ненулевые. Отметим, что мы не требуем, чтобы $\det(A_{nn})$ (т. е. $\det A_0$) был отличен от нуля.

Предположим, что первый равный нулю ведущий элемент есть $(r+1)$ -й, так что первые r главных миноров ненулевые, тогда как $\det(A_{r+1, r+1})$ равен нулю. Из (19.6), если $\det(A_{i, r+1}) \neq 0$ для некоторого i , большего $r+1$, следует, что алгоритм не может осуществиться, потому что тогда $n_{i, r+1}$ равно бесконечности. Однако если $\det(A_{i, r+1}) = 0$ ($i = r+1, \dots, n$), то все $n_{i, r+1}$ не определены. В этом случае можем взять для $n_{i, r+1}$ произвольные значения, и если, в частности, возьмем их равными нулю, то $N_{r+1} = I$ и $(r+1)$ -й основной шаг оставляет A_r неизменной.

Действительно, в этом случае первый столбец $W_{n-r, n-r}$ нулевой, что можно увидеть из (19.14), если мы заметим, что

$$A_{rr, i, r+1} = A_{i, r+1}. \quad (20.1)$$

Далее, ранг первых $(r+1)$ столбцов A_r , очевидно, равен r , что показывает их линейную зависимость. Однако эти столбцы получены из соответствующих столбцов A_0 с помощью левого умножения на невырожденные матрицы и, следовательно, они также должны быть линейно зависимы. Поэтому A_0 — вырожденная, и решение системы $Ax = b$ существует для таких b , при которых ранг A_0 равен рангу $(A_0 | b)$. Если уравнения несовместны, то это обнаруживается в обратной подстановке на том шаге, когда мы пытаемся вычислить x_{r+1} делением на нулевой ведущий элемент. Если они совместны, то x_{r+1} будет неопределенно.

Заметим, что несостоятельность процесса исключения, вообще говоря, не связана с вырожденностью или плохой обусловленностью матрицы. Например, если A_0 такова:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

то метод Гаусса неосуществим уже на первом шаге, хотя A_0 — ортогональная и поэтому хорошо обусловлена.

Численная устойчивость

21. Пока мы рассмотрели лишь возможность несостоятельности процесса исключения. Однако если посмотреть теперь на этот факт с точки зрения вычислений, то процесс, вероятно, должен быть численно неустойчив, если какой-либо из множителей $n_{i, r}$ велик. Чтобы преодолеть это, мы можем заменить N_r на устойчивую матрицу N_r , описанную в гл. 3, § 47. На r -м основном шаге мы умножаем на $N_r I_r$, где r' определяется соотношением

$$|a_{r', r}^{(r-1)}| = \max_{i=r, \dots, n} |a_{ir}^{(r-1)}|. \quad (21.1)$$

Если r' определяется таким образом не единственно, мы берем наименьшее из его возможных значений.

Левое умножение на $I_{r,r'}$ приводит к *перестановке* строк r и r' (где $r' \geq r$), а последующее левое умножение на N_r — к вычитанию кратного новой r -й строки по очереди из каждой строки от $(r+1)$ -й до n -й, причем все множители ограничены по модулю единицей. Отметим, что строки с 1-й до $(r-1)$ -й остаются на r -м шаге без изменения, и сохраняются нули, полученные на предыдущих шагах. Для модифицированной процедуры имеем

$$A_r = N_r I_{r,r'} \dots N_2 I_{2,2'} N_1 I_{1,1'} A_0, \quad (21.2)$$

где A_r по-прежнему вида (18.2).

Устойчивый метод единственным образом определяется на каждом шаге, если только в начале r -го шага не все $a_{ir}^{(r-1)}$ ($i = r, \dots, n$) равны нулю. Тогда наше правило дает $r' = r$, но множители не определены. Мы можем взять $N_r = I$, и в этом случае r -й шаг пропускается, что допустимо, так как A_{r-1} — уже треугольная в первых r столбцах. Как и в неустойчивом методе, это может случиться лишь тогда, когда первые r столбцов A_0 линейно зависимы. Устойчивый процесс исключения обычно называется *методом Гаусса с выбором главного элемента по столбцу*. Заметим, что он осуществляется при всех обстоятельствах, а при условиях, которые мы описали, к тому же единственным образом. Если уравнения несовместны, то это станет ясно при обратной подстановке.

Значение перестановок

22. Описание модифицированного процесса становится более простым, если мы предполагаем фактическое выполнение перестановок строк, так как тогда окончательная матрица A_{n-1} будет точно верхней треугольной и все промежуточные матрицы — вида (15.6). На некоторых вычислительных машинах может быть более удобно оставлять уравнения на своих прежних местах и сохранять информацию о номерах ведущих строк. Тогда окончательная матрица такая же, как, например, для случая $n = 5$:

$$\begin{bmatrix} 0 & 0 & 0 & \times & \times \\ 0 & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times \end{bmatrix}, \quad (22.1)$$

где последовательными ведущими строками были третья, вторая, четвертая, первая и пятая. Эта матрица треугольная с точностью до перестановки строк.

Если мы действительно делаем перестановки, то

$$A_{n-1}x - b_{n-1} = N_{n-1}I_{n-1,(n-1)'} \dots N_2 I_{2,2'} N_1 I_{1,1'} (A_0x - b_0). \quad (22.2)$$

Отсюда видно, что в действительности выбор главного элемента определяет такую перестановку строк A_0 , при которой метод Гаусса может быть выполнен и без выбора главного элемента, причем все множители будут ограничены по модулю единицей.

Для того чтобы избежать усложнения доказательства, мы в качестве иллюстрации рассмотрим случай $n = 4$. Из (22.2) следует, что

$$\begin{aligned} A_3 &= N_3 I_{3,3'} N_2 I_{2,2'} N_1 I_{1,1'} A_0 = \\ &= N_3 I_{3,3'} N_2 (I_{3,3'} I_{3,3'}) I_{2,2'} N_1 (I_{2,2'} I_{3,3'} I_{3,3'} I_{2,2'}) I_{1,1'} A_0, \end{aligned} \quad (22.3)$$

так как произведения в скобках равны единичной матрице. Перегруппируя члены, имеем

$$A_3 = N_3(I_{3,3} \cdot N_2 I_{3,3'}) (I_{3,3} I_{2,2} \cdot N_1 I_{2,2'} I_{3,3'}) (I_{3,3} I_{2,2} I_{1,1} \cdot A_0) = \\ = N_3 \tilde{N}_2 \tilde{N}_1 \tilde{A}_0, \quad (22.4)$$

где, согласно гл. 1, § 41 (iv), $\tilde{N}_2$ и $\tilde{N}_1$ получаются из N_2 и N_1 лишь перестановкой поддиагональных элементов, а $\tilde{A}_0$ из A_0 — перестановкой строк. Предположим теперь, что мы должны выполнить метод Гаусса для $\tilde{A}_0$ без выбора главного элемента. Тогда мы получили бы

$$A_3^* = N_3^* N_2^* N_1^* \tilde{A}_0, \quad (22.5)$$

где A_3^* — верхняя треугольная, а N_3^* , N_2^* , N_1^* — элементарные матрицы обычного типа. Уравнения (22.4) и (22.5) дают два треугольных разложения $\tilde{A}_0$ (с нижними треугольными матрицами, имеющими единичные диагональные элементы) и, следовательно, они совпадают.

Это означает, что $\tilde{N}_3 = N_3^*$, $\tilde{N}_2 = N_2^*$, $\tilde{N}_1 = N_1^*$, и поэтому метод Гаусса для $\tilde{A}_0$ без перестановок приводит к множителям, ограниченным по модулю единицей.

Численный пример

23. В табл. 1 мы показываем применение метода Гаусса с перестановками на примере решения системы уравнений четвертого порядка в 4-значной десятичной арифметике с фиксированной запятой. Работая с карандашом и бумагой, неудобно выполнять перестановки, так как это приводит к переписыванию уравнений в нужном порядке. Вместо этого мы подчеркнули ведущие уравнения. Мы включили уравнения, которые были использованы как ведущие, в каждую из преобразованных систем, так что все эти системы эквивалентны исходной. Решение вычислено с одним и тем же числом разрядов после запятой, и самая большая его компонента имеет 4 знака. Другими словами, мы получили 4-значное блочноплавающее решение.

Сразу же можно видеть, что решение вряд ли должно иметь больше двух верных знаков в наибольшей компоненте, так как она получается делением 0,5453 на — 0,0052. Присутствие в делителе лишь одной ошибки округления будет означать, что и без накопления ошибок мы должны быть готовы к ошибке во втором знаке в вычисленном x_4 . В общем случае, если в каком-либо ведущем элементе потеряно k верных знаков, нам следует ожидать соответствующей потери и в некоторых компонентах решения. В табл. 1 мы даем ошибки округления, полученные на каждом шаге преобразования. Значение этих ошибок обсуждается в §§ 24, 25. Мы также приводим N и A_3 , переставленные таким образом, чтобы дать правильные треугольные матрицы, и соответственно переставленное точное произведение $N(A_3 : b_3)$. При точном вычислении в триангуляризации оно должно быть равно $(A_0 : b_0)$ со строками 4, 2, 3, 1 в порядке 1, 2, 3, 4. Заметим, что матрица $N(A_3 : b_3)$ действительно близка к переставленной, причем наибольшее отклонение равно 0,00006248 в элементе (1,4).

Наконец, мы даем точные значения $A_0 x$ для вычисленного x и вектора r , определенного равенством $r = b_0 - A_0 x$. Этот вектор называется *вектором-невязкой*, соответствующим x , а его компоненты называются *невязками*. Можно видеть, что максимальная невязка равна 0,03296. Это необычайно мало, так как мы ожидаем ошибку порядка единицы

Таблица 1

n	x_1	x_2	x_3	x_4	b	$10^4 \epsilon_i^{(1)}$	$10^4 \epsilon_i^{(2)}$	$10^4 \epsilon_i^{(3)}$	r (точное)
$A_0 x = b_0$									
0,2678)	0,2317	0,6123	0,4137	0,6696	0,4753	0,2734	0,3670	0,1870	0,03296
0,4950)	0,4283	0,8176	0,4257	0,8312	0,2167	0,2350	0,1750	-0,3250	0,02703
0,8461)	0,7321	0,4135	0,3126	0,5163	0,8132	0,3033	0,0165	-0,1150	0,00317
	0,8653	0,2165	0,8265	0,7123	0,5165	0	0	-0,2297	0,01005
$A_1 x = b_1$									
0,7803)	0,0000	0,5543	0,1924	0,4788	0,3370	0	-0,4702	-0,4842	
	0,0000	0,7104	0,0166	0,4786	-0,0390	0	0	0	
0,3242)	0,0000	0,2303	-0,3867	-0,0864	0,3762	0	-0,1828	-0,3788	
	0,8653	0,2165	0,8265	0,7123	0,5165	0	0	0	
$A_2 x = b_2$									
-0,4575)	0,0000	0,0000	0,1794	0,1053	0,3674	0	-0,1425	0,3200	
	0,0000	0,7104	0,0166	0,4786	-0,0390	0	0	0	
	0,0000	0,0000	-0,3921	-0,2416	0,3888	0	0	0	
	0,8653	0,2165	0,8265	0,7123	0,5165	0	0	0	
$A_3 x = b_3$									
	0,0000	0,0000	0,0000	-0,0052	0,5453	8,9	0,44234		
	0,0000	0,7104	0,0166	0,4786	-0,0390	69,1	0,18967		
	0,0000	0,0000	-0,3921	-0,2416	0,3888	63,6	0,81003		
	0,8653	0,2165	0,8265	0,7123	0,5165	-104,9	0,50645		
$(A_3; b_3)$ (переставленные)									
	1,0000	1,0000	1,0000	1,0000	0,2165	0,8265	0,7123	0,5165	
	0,4950	0,3242	1,0000	1,0000	0,7104	0,0166	0,4786	-0,0390	
	0,8461	0,7803	-0,4575	1,0000	0,8653	0,7104	0,0166	0,3888	
	-0,2678	0,7803	-0,4575	1,0000	-0,3921	-0,2416	-0,0052	0,5453	
$N(A_3; b_3)$ (точное) (ср. с $(A_0; b_0)$ с переставленными строками)									
	0,8653	0,0000	0,2165	0,0000	0,7123	0,0000	0,8265	0,0000	
	0,4283	2350	0,8175	6750	0,4257	1750	0,8311	8850	
	0,7321	3033	0,4134	9233	0,3125	8337	0,5162	3915	
	0,2317	2734	0,6123	0382	0,4136	7543	0,6695	3752	
								0,4753	
								1100	

имеем для $j \geq i$:

$$n_{i1}a_{1j}^{(0)} + n_{i2}a_{2j}^{(1)} + \dots + n_{i, i-1}a_{i-1, j}^{(i-2)} + a_{ij}^{(i-1)} = a_{ij}^{(0)} + \varepsilon_{ij}^{(1)} + \varepsilon_{ij}^{(2)} + \dots + \varepsilon_{ij}^{(i-1)}, \quad (24.5)$$

а для $i > j$ с включенным уравнением (24.4)

$$n_{i1}a_{1j}^{(0)} + n_{i2}a_{2j}^{(1)} + \dots + n_{ij}a_{jj}^{(j-1)} = a_{ij}^{(0)} + \varepsilon_{ij}^{(1)} + \varepsilon_{ij}^{(2)} + \dots + \varepsilon_{ij}^{(j)}. \quad (24.6)$$

Аналогично для правой части системы имеем

$$b_i^{(k)} = b_i^{(k-1)} - n_{ik}b_k^{(k-1)} + \varepsilon_i^{(k)} \quad (k=1, \dots, i-1), \quad (24.7)$$

откуда

$$n_{i1}b_1^{(0)} + n_{i2}b_2^{(1)} + \dots + n_{i, i-1}b_{i-1}^{(i-2)} + b_i^{(i-1)} = b_i^{(0)} + \varepsilon_i^{(1)} + \dots + \varepsilon_i^{(i-1)}. \quad (24.8)$$

Уравнения (24.5), (24.6) и (24.8) выражают тот факт, что

$$N(A_{n-1} \vdots b_{n-1}) = (A_0 + F \vdots b_0 + k),$$

где

$$\left. \begin{aligned} f_{ij} &= \varepsilon_{ij}^{(1)} + \dots + \varepsilon_{ij}^{(i-1)} & (j \geq i), \\ f_{ij} &= \varepsilon_{ij}^{(1)} + \dots + \varepsilon_{ij}^{(j)} & (i > j), \\ k_i &= \varepsilon_i^{(1)} + \dots + \varepsilon_i^{(i-1)}. \end{aligned} \right\} \quad (24.9)$$

Другими словами, вычисленные матрицы N и A_{n-1} являются матрицами, которые могли быть получены при точном разложении $(A + F)$. Это точный результат, так как нет никакого пренебрежения членами второго порядка малости. Данный результат согласуется с нашим анализом главы 3, но теперь мы находимся в лучшем положении, чтобы получить верхние оценки для ошибок округления.

Выделим два главных следствия из этих результатов:

I. Точный определитель A_{n-1} , который равен

$$a_{11}^{(0)} a_{22}^{(1)} \dots a_{nn}^{(n-1)},$$

есть точный определитель $(A_0 + F)$.

II. Точное решение системы $A_{n-1}x = b_{n-1}$ есть точное решение системы $(A + F)x = b_0 + k$.

Верхние оценки матриц возмущения для арифметики с фиксированной запятой

25. Предположим, что исходные уравнения были нормированы и уравновешены по крайней мере приближенно и были сделаны перестановки. Предположим далее, что все $a_{ij}^{(k)}$ и $b_i^{(k)}$ ограничены по модулю единицей. В общем случае $\varepsilon_{ij}^{(k)}$ есть ошибка, сделанная при вычислении $a_{ij}^{(k)}$ из вычисленных значений $a_{ij}^{(k-1)}$, n_{ik} и $a_{kj}^{(k-1)}$. В этом процессе сделана лишь одна ошибка и то в тот момент, когда $2t$ — разрядное произведение $n_{ik}a_{kj}^{(k-1)}$ округляется до t разрядов. Следовательно,

$$|\varepsilon_{ij}^{(k)}| \leq \frac{1}{2} \cdot 2^{-t}. \quad (25.1)$$

Это справедливо для всех $\varepsilon_{ij}^{(k)}$, за исключением тех из них, которые возникают из-за ошибки округления при делении во время вычислений n_{ij} .

Из (24.3) следует, что ξ_{ij} удовлетворяет неравенству

$$|\xi_{ij}| \leq \frac{1}{2} \cdot 2^{-t}, \quad (25.2)$$

так как это ошибка одного деления; мы уверены, что $|n_{ij}| \leq 1$, так как $|a_{ij}^{(j-1)}| \geq |a_{ij}^{(j-1)}|$, и поэтому округление не может привести к значению, которое по модулю больше единицы. Следовательно, из (24.4) и нашего предположения относительно элементов преобразованных матриц имеем

$$|\varepsilon_{ij}^{(j)}| = |a_{ij}^{(j-1)} \xi_{ij}| \leq \frac{1}{2} \cdot 2^{-t}. \quad (25.3)$$

Аналогично при сделанном предположении относительно элементов $b_i^{(k)}$

$$|\varepsilon_i^{(k)}| \leq \frac{1}{2} \cdot 2^{-t}. \quad (25.4)$$

Объединяя соотношения (25.1), (25.3) и (25.4) для всех соответствующих k , получаем

$$|k_i| \leq \frac{1}{2} (i-1) \cdot 2^{-t}, \quad |f_{ij}| \leq \begin{cases} \frac{1}{2} (i-1) 2^{-t} & (i \leq \bar{i}), \\ \frac{1}{2} j 2^{-t} & (i > \bar{i}), \end{cases} \quad (25.5)$$

причем этот результат дает максимальные верхние оценки.

Ошибки на каждом шаге исключения для предыдущего примера представлены в табл. 1. Внизу таблицы показана точная матрица $N (A_3 : b_3)$. Ее следует сравнить с матрицей $(A_0 : b_0)$ с соответственно переставленными строками. Заметим, что разность между ними есть точная сумма матриц ошибок на каждом шаге.

Верхняя оценка для элементов преобразованных матриц

26. Пока предположение, что $|a_{ij}^{(k)}| \leq 1$ для всех i, j, k , должно казаться совсем необоснованным. Даже с выбором главного элемента мы можем гарантировать лишь то, что $|a_{ij}^{(k)}| \leq 2^k$. Это следует из таких неравенств:

$$|a_{ij}^{(k)}| = |a_{ij}^{(k-1)} - n_{ik} a_{kj}^{(k-1)}| \leq |a_{ij}^{(k-1)}| + |a_{kj}^{(k-1)}|. \quad (26.1)$$

К сожалению, мы можем построить такие матрицы (по общему признанию весьма искусственные), для которых эта граница достигается. Например, матрицу вида

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}. \quad (26.2)$$

Выбор главного элемента по всей матрице

27. Наш анализ уже показал особую важность предохранения последовательных элементов $a_{ij}^{(k)}$ от быстрого роста. Теперь рассмотрим другой способ выбора главного элемента, для которого можно дать значительно меньшие оценки $|a_{ij}^{(k)}|$.

В алгоритме, описанном в § 21, неизвестные исключаются в естественном порядке, и на r -м шаге главный элемент выбирается в первом столбце квадратной матрицы порядка $(n - r + 1)$, находящейся в нижнем правом углу A_{r-1} .

Предположим теперь, что мы выбираем в качестве r -го ведущего элемента максимальный по модулю элемент во всей матрице. Если этот элемент находится в позиции (r', r'') , где $r', r'' \geq r$, то, выполняя перестановки, можем написать

$$A_r = N_r I_{r, r'} A_{r-1} I_{r, r''}. \quad (27.1)$$

В общем случае требуются перестановки как строки, так и столбца, и неизвестные больше не исключаются в естественном порядке, однако окончательная матрица A_{n-1} все еще верхняя треугольная. На примере матрицы четвертого порядка имеем

$$(N_3 I_{3,3} \cdot N_2 I_{2,2} \cdot N_1 I_{1,1} \cdot A_0 I_{1,1} \cdot I_{2,2} \cdot I_{3,3}) (I_{3,3} \cdot I_{2,2} \cdot I_{1,1} \cdot x) = \\ = N_3 I_{3,3} \cdot N_2 I_{2,2} \cdot N_1 I_{1,1} \cdot b_0, \quad (27.2)$$

так как произведение множителей между A_0 и x равно единичной матрице. Поэтому решение $A_r y = b_r$ есть $I_{r, r''} \dots I_{2,2''} I_{1,1''} x$.

Здесь мы используем двухсторонние эквивалентные преобразования A_0 , но правые множители суть просто матрицы перестановок.

Назовем эту модификацию *выбором главного элемента по всей матрице*, в отличие от прежнего процесса, который назовем *выбором главного элемента по столбцу*. Уилкинсон (1961b) показал, что в этом случае

$$|a_{ij}^{(r-1)}| \leq r^{1/2} [2^{1/2} 4^{1/3} \dots r^{1/(r-1)}]^{1/2} a = f(r) a, \quad (27.3)$$

где $a = \max |a_{ij}^0|$. В табл. 2 даны значения $f(r)$ для некоторых значений r .

Т а б л и ц а 2

r	10	20	50	100
$f(r)$	19	67	530	3300

Функция $f(r)$ много меньше, чем 2^r для больших r , а метод доказательства (27.3) показывает, что достижимая верхняя оценка существенно меньше. В действительности до сих пор не найдено ни одной матрицы, для которой $f(r) > r$. Для матриц вида (26.2) любого порядка максимальный элемент при выборе главного элемента по всей матрице равен 2.

Невольно можно заключить, что было бы целесообразно использовать выбор главного элемента по всей матрице во всех случаях, однако выбор главного элемента по столбцу имеет существенные преимущества. Наиболее важные из них следующие.

(I) На вычислительных машинах с двухступенчатой памятью легче организовать выбор главного элемента по столбцу, чем по всей матрице.

(II) Если матрица имеет значительное число нулевых элементов, то некоторые структуры их расположения сохраняются при выборе

главного элемента по столбцу и разрушаются при выборе главного элемента по всей матрице. С нашей точки зрения, это очень убедительный аргумент.

Несмотря на существование матриц вида (26.2), мы думаем, что любое значительное увеличение элементов A_r нетипично даже при выборе главного элемента по столбцу. Если матрица A_0 вообще плохо обусловлена, то для элементов A_r , как правило, имеет место устойчивая тенденция к *уменьшению*, особенно если элемент (i, j) есть гладкая функция i и j , так как в этом случае процесс исключения дает результаты, близкие к вычислению разностей. В моей практике не было ни одного естественного примера, в котором элементы увеличивались бы более, чем в 16 раз.

Практический процесс с выбором главного элемента по столбцу

28. Рекомендуемый процесс с выбором главного элемента по столбцу в арифметике с фиксированной запятой состоит в следующем. Сначала нормируем матрицу так, чтобы все элементы лежали в пределах от $-1/2$ до $+1/2$. Поэтому мы можем выполнить первый шаг без опасения, что элементы превзойдут единицу. После того как выполнен этот шаг, проверяем каждый вычисленный элемент, чтобы посмотреть, не превзошел ли его модуль $1/2$. Если какой-либо элемент стал по модулю больше $1/2$, то вся соответствующая строка делится на 2. Итак, начинаем следующий шаг с матрицы, элементы которой снова лежат в пределах от $-1/2$ до $+1/2$.

Соответствующий анализ ошибок требует лишь небольшого изменения. При делении строки на 2 в каждый элемент дополнительно вносится ошибка округления, ограниченная величиной 2^{-t} . Последующие ошибки округления, внесенные в эту строку, эквивалентны возмущениям в исходной матрице, ограниченным величиной 2^{-t} , а не $\frac{1}{2} \cdot 2^{-t}$, при этом сами ошибки все еще аддитивны. Аналогичное замечание относится и к тому случаю, когда строка делится более одного раза.

Анализ ошибок с плавающей запятой

29. Процесс с фиксированной запятой, описанный в § 28, в некотором отношении более похож на процесс с плавающей запятой, хотя мы и отмечали, что на практике деление строк должно осуществляться редко. Анализ ошибок вычислений с настоящей плавающей запятой во многом следует той же самой линии. Здесь имеем

$$\begin{aligned} a_{ij}^{(k)} &= fl(a_{ij}^{(k-1)} - n_{ik}a_{kj}^{(k-1)}) \equiv \\ &\equiv [a_{ij}^{(k-1)} - n_{ik}a_{kj}^{(k-1)}(1 + \varepsilon_1)](1 + \varepsilon_2) \quad (|\varepsilon_i| \leq 2^{-t}), \end{aligned} \quad (29.1)$$

$$n_{ij} = fl(a_{ij}^{(j-1)}/a_{jj}^{(j-1)}) \equiv a_{ij}^{(j-1)}/a_{jj}^{(j-1)} + \xi_{ij} \quad (|\xi_{ij}| \leq \frac{1}{2} \cdot 2^{-t}), \quad (29.2)$$

где можем пользоваться абсолютной оценкой для ξ_{ij} , так как выбор главного элемента обеспечивает условие $|n_{ij}| \leq 1$. Если мы предположим, что

$$|a_{ij}^{(k)}| \leq a_k, \quad (29.3)$$

то при тех же обозначениях, что и в анализе с фиксированной запятой, имеем

$$|\varepsilon_{ij}^{(j)}| = |a_{jj}^{(j-1)} \xi_{ij}| \leq 2^{-t-1} a_{j-1} \quad (i > j), \quad (29.4)$$

а для всех других i, j и k простые вычисления показывают, что

$$|\varepsilon_{ij}^{(k)}| = \left| \frac{a_{ij}^{(k)} \varepsilon_2}{1 + \varepsilon_2} - n_{ik} a_{kj}^{(k-1)} \varepsilon_1 \right| \leq \leq 2^{-t} [a_k / (1 - 2^{-t}) + a_{k-1}] \leq 2^{-t} [(1,01) a_k + a_{k-1}]. \quad (29.5)$$

Эти оценки показывают, что вычисленные N и A_{n-1} таковы, что

$$NA_{n-1} = A_0 + F, \quad (29.6)$$

где

$$|f_{ij}| \leq \begin{cases} 2^{-t} [a_0 + (2,01) a_1 + \dots + (2,01) a_{i-2} + (1,01) a_{i-1}] & (j \geq i), \\ 2^{-t} [a_0 + (2,01) a_1 + \dots + (2,01) a_{j-2} + (1,51) a_{j-1}] & (i > j). \end{cases} \quad (29.7)$$

Оценки (29.7) могут быть несколько уточнены, но это несущественно. Эти оценки показывают, что если $|a_{ij}^{(h)}| \leq a$ для всех соответствующих i, j, k , то

$$|f_{ij}| \leq \begin{cases} (2,01) 2^{-t} a (i-1), \\ (2,01) 2^{-t} a j. \end{cases} \quad (29.8)$$

Из уравнений (29.7) следует, что когда $a_{ij}^{(h)}$ постепенно уменьшаются с ростом k , то оценка F для плавающей запятой может быть много лучше, чем для фиксированной. Если, например,

$$a_k < 2^{-k}, \quad (29.9)$$

так что на каждом шаге теряется приблизительно один двоичный знак, то

$$|f_{ij}| \leq 2^{-t} [1 + (2,01) (2^{-1} + 2^{-2} + \dots)] < (3,01) 2^{-t}. \quad (29.10)$$

Постепенная потеря значащих цифр — довольно общее явление для плохо обусловленных уравнений, и наш результат показывает, что вычисление с плавающей запятой может быть лучше; это подтверждалось на практике. (Аналогичный эффект может быть получен и с фиксированной запятой, если после каждого преобразования будем умножать каждую строку на наибольшую степень 2, для которой все элементы остаются в пределах ± 1 .)

Мы можем получить сравнимые результаты для ошибок округления, сделанных в правой части системы.

Разложение с плавающей запятой без выбора главного элемента

30. В табл. 3 мы приводим простой пример, иллюстрирующий катастрофическую потерю точности, которая может стать следствием недостаточной величины ведущего элемента. Рассматриваемая матрица очень хорошо обусловлена, но малая величина ведущего элемента неизбежно влияет на точность. Использовалось вычисление с плавающей запятой, так как элементы матриц A , имеют большой разброс.

Т а б л и ц а 3

n	A_0			b_0
	0,000003	0,213472	0,332147	0,235262
71837,3)	0,215512	0,375623	0,476625	0,127653
57752,3)	0,173257	0,663257	0,625675	0,285321
	A_1			b_1
	0,000003	0,213472	0,332147	0,235262
	0,000000	-15334,9	-23860,0	-16900,5
0,803905)	0,000000	-12327,8	-19181,7	-13586,6
	A_2			b_2
	0,000003	0,213472	0,332147	0,235262
	0,000000	-15334,9	-23860,0	-16900,5
	0,000000	0,000000	-0,50000	-0,200000

Уравнения с эквивалентным возмущением

0,000003	0,213472	0,332147	0,235262
0,2155119	0,3521056	9,5436831	0,0868726
0,1732569	0,6989856	0,5531881	0,3216026

Первый ведущий элемент очень мал, и в результате этого n_{21} и n_{31} будут порядка 10^5 . Соответственно и вычисленные элементы A_1 и b_1 также очень велики. Если мы рассмотрим вычисление $a_{23}^{(1)}$, то получим

$$a_{23}^{(1)} = fl(a_{23}^{(0)} - n_{21}a_{13}^{(0)}) = fl(0,476625 - 71837,3 \times 0,332147) = \\ = fl(0,476625 - 23860,5) = -23860,0.$$

Точное значение $a_{23}^{(0)} - n_{21}a_{13}^{(0)}$ таково: $0,476625 - 23860,5436831 = -23860,0670581$, и величина $\epsilon_{23}^{(1)}$ есть разность между вычисленным и точным значениями. Заметим, что при вычислении $a_{23}^{(1)}$ не учитываются почти все знаки в $a_{23}^{(0)}$; мы получили бы то же самое численное значение $a_{23}^{(1)}$ для любого элемента $a_{23}^{(0)}$, лежащего в пределах от 0,45 до 0,549999.

В табл. 3 даны уравнения с эквивалентным возмущением. Отсюда видно, что возмущенные второе и третье уравнения не имеют никакого отношения к исходным. Легко проверить, что решение этой системы уравнений сильно отличается от решения исходной системы.

Мы показали также второй основной шаг вычислений. Заметим, что большие элементы, которые появлялись в третьем уравнении в результате первого шага, исчезают на втором шаге. Однако очевидно, что почти все знаки, приведенные в вычисленном $a_{33}^{(2)}$, неверные. Значение определителя, данное произведением $a_{11}^{(0)}a_{22}^{(1)}a_{33}^{(2)}$, также очень неточно, так как $a_{33}^{(2)}$ имеет самое большое один верный знак.

Потеря верных знаков

31. До сих пор мы сосредоточивали внимание на опасности применения тех методов, которые приводят к увеличению элементов матриц A . Однако можно возразить, что более серьезной является потеря точности в тех примерах, где имеет место постепенное уменьшение величины элементов. Это действительно так, но между этими двумя явлениями существует и важное различие. Если мы не выбираем главный элемент или

выбираем его лишь по столбцу, а элементы преобразованной матрицы очень велики, то можем потерять точность, но это, вообще говоря, не обязательно и может быть устранено другим выбором ведущих элементов.

С другой стороны, постепенное уменьшение верных знаков есть следствие плохой обусловленности, и потеря точности более или менее неизбежна. Доказательство § 29 показывает, что если используется плавающая запятая, то при потере верных знаков во время процесса исключения эквивалентное возмущение в исходной матрице фактически меньше, чем в том случае, когда потери знаков нет.

Общераспространенная ошибка

32. Существует широко распространенное убеждение, что выбор больших ведущих элементов в некотором отношении вреден. Аргумент обычно следующий. Произведение ведущих элементов равно определителю A_0 и поэтому не зависит от их выбора. Если мы вначале выбираем большие ведущие элементы, то в конце они обязательно должны быть меньше и поэтому будут иметь большие относительные ошибки.

Итак, существуют матрицы, для которых последний ведущий элемент мал при выборе главного элемента во всей матрице, но не при выборе главного элемента в столбце. Однако этот ведущий элемент есть обратная величина некоторого элемента промежуточной обратной матрицы и, следовательно, не может быть меньше чем 2^{-k} , если только норма обратной матрицы не больше 2^k . Если матрица в действительности имеет столь большую обратную матрицу, то соответственно падает и точность. Появление малого ведущего элемента есть просто один из видов проявления плохой обусловленности. В примере из § 30 мы видели, что использование не являющегося необходимым малого ведущего элемента может и не сделать последний ведущий элемент большим; *это приводит лишь к очень большому ведущему элементу на следующем шаге и возвращению к «нормальному» на последующих.*

Мы не утверждаем, что выбор наибольшего элемента в качестве ведущего является наилучшим, и в действительности это часто будет неверно. Мы лишь утверждаем, что до сих пор не предложен никакой другой практически приемлемый процесс. Одно из типичных предложений заключается в том, что в качестве ведущего элемента следует выбирать элемент, ближайший по модулю к корню n -й степени из абсолютного значения определителя. Есть три возражения против этого и аналогичных ему предложений:

(i) Значение определителя обычно неизвестно до тех пор, пока не выполнено исключение.

(ii) Может не быть элементов, близких к предписанному значению.

(iii) Даже если (i) и (ii) неверны, нетрудно построить примеры, для которых этот способ неудачен. С другой стороны, выбор главного элемента по всей матрице никогда не является очень плохим.

Матрицы специального вида

33. Есть три специальных класса матриц, которые имеют большое значение в решении проблемы собственных значений и для которых выбор ведущего элемента осуществляется много проще. Доказательство следующих результатов дано очень кратко, так как оно просто и детально разобрано Уилкинсоном (1961b).

Класс I. Матрицы Хессенберга. Мы назовем матрицу A *верхней матрицей Хессенберга*, если

$$a_{ij} = 0 \quad (i \geq j + 2). \quad (33.1)$$

Аналогично будем говорить, что A есть *нижняя матрица Хессенберга*, если

$$a_{ij} = 0 \quad (j \geq i + 2), \quad (33.2)$$

так что, если A — верхняя матрица Хессенберга, то A^T — нижняя. Если мы выполняем исключение Гаусса с выбором главного элемента по столбцу для верхней матрицы Хессенберга, то перед началом r -го основного шага A_{r-1} будет такого же вида, как матрица .

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ & 0 & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{bmatrix}, \quad (33.3)$$

соответствующая $n = 6$, $r = 3$. Здесь мы выделили нули, полученные на предыдущих шагах. Матрица, стоящая в нижнем правом углу, сама является верхней матрицей Хессенберга. Элементы в строках от $(r + 1)$ -й до n -й те же, что и в A_0 . Есть лишь две строки в соответствующей матрице, в которые входят x_r , так что r' равно или r , или $r + 1$. Можно сказать, что в зависимости от того, $r' = r + 1$ или $r' = r$, перестановка на r -м шаге имеет место или нет. Из вышесказанного вытекает, что если

$$|a_{ij}^{(0)}| \leq 1, \quad \text{то} \quad |a_{ij}^{(r)}| \leq r + 1, \quad (33.4)$$

и каждый столбец $\tilde{N}$ (переставленной N) имеет лишь один поддиагональный элемент. Если, например, для матрицы шестого порядка перестановки имеют место на шагах 1, 3, 4, то $\tilde{N}$ такова:

$$\begin{bmatrix} 1 & & & & & \\ \times & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & \times & \times & \times & 1 & \\ 0 & 0 & 0 & 0 & \times & 1 \end{bmatrix}. \quad (33.5)$$

Если исключение Гаусса требуется выполнить для нижней матрицы Хессенберга, то неизвестные следует исключить в порядке $x_n, x_{n-1}, \dots, x_1$, так что окончательная матрица будет нижней треугольной. В любом случае использование выбора главного элемента по всей матрице разрушает простую структуру элементов, а так как для $a_{ij}^{(r)}$ имеем теперь существенно лучшую оценку, то в этом процессе нет особой необходимости.

Класс II. Трехдиагональные матрицы. Мы определили трехдиагональные матрицы в гл. 3, § 13; из этого определения вытекает, что трехдиагональные матрицы являются одновременно как верхними, так и нижними матрицами Хессенберга. Поэтому результаты, которые мы получили для верхних матриц Хессенберга, справедливы и для трехдиагональных матриц, но допускают дальнейшее упрощение. Здесь $\tilde{N}$ имеет тот же вид,

что и прежде, тогда как A_0 и A_r таковы ($n = 6$, $r = 3$):

$$A_0 = \begin{bmatrix} \alpha_1 & \beta_1 & & & & \\ \gamma_2 & \alpha_2 & \beta_2 & & & \\ & \gamma_3 & \alpha_3 & \beta_3 & & \\ & & \gamma_4 & \alpha_4 & \beta_4 & \\ & & & \gamma_5 & \alpha_5 & \beta_5 \\ & & & & \gamma_6 & \alpha_6 \end{bmatrix}, \quad A_3 = \begin{bmatrix} u_1 & v_1 & w_1 & & & \\ & u_2 & v_2 & w_2 & & \\ & & u_3 & v_3 & w_3 & \\ & & & p_4 & q_4 & \\ & & & & \gamma_5 & \alpha_5 & \beta_5 \\ & & & & & \gamma_6 & \alpha_6 \end{bmatrix}. \quad (33.6)$$

Элемент w_i равен нулю, если только не было перестановки на i -м шаге. Из простого индуктивного доказательства следует, что если

$$|\alpha_i|, |\beta_i|, |\gamma_i| \leq 1 \quad (i = 1, \dots, n), \quad (33.7)$$

то

$$|p_i| \leq 2, |q_i| \leq 1, |u_i| \leq 2, |v_i| \leq 1, |w_i| \leq 1. \quad (33.8)$$

Класс III. Положительно определенные симметричные матрицы. Для матриц этого типа исключение Гаусса вполне устойчиво без какого-либо специального выбора ведущих элементов. Мы можем показать, что если, как и в § 18,

$$A_r = \left[\begin{array}{c|c} U_{rr} & U_{r, n-r} \\ \hline 0 & W_{n-r, n-r} \end{array} \right], \quad (33.9)$$

то $W_{n-r, n-r}$ сама положительно определенная и симметричная на каждом шаге, и если $|a_{ij}^{(0)}| \leq 1$, то $|a_{ij}^{(k)}| \leq 1$; (смотри, например, Уилкинсон (1961b)). Опускаем доказательство, так как хотя триангуляризация положительно определенных матриц чрезвычайно важна для наших приложений, мы не будем для этого использовать исключение Гаусса.

Заметим, что мы не утверждаем, что все элементы N ограничены единицей, и в действительности это может быть неверно. В качестве иллюстрации рассмотрим матрицу

$$\begin{bmatrix} 0,000064 & 0,006873 \\ 0,006873 & 1,000000 \end{bmatrix}, \quad (33.10)$$

для которой $n_{21} = 107,391$. Это означает, что мы должны использовать арифметику с плавающей запятой, однако большие множители не могут привести к увеличению максимального из элементов преобразованных матриц. В нашем примере $a_{22}^{(1)} = 0,261902$. Наш анализ показал, что эквивалентное возмущение исходной матрицы на каждом шаге не больше, чем сами ошибки. Так как все элементы всех преобразованных матриц ограничены единицей, ошибки округления не могут быть большими.

Метод Гаусса на быстродействующей вычислительной машине

34. Теперь мы опишем организацию вычислений по методу Гаусса с выбором главного элемента по столбцу на быстродействующей машине. В начале r -го шага состояние памяти аналогично случаю $n = 5$, $r = 3$:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & b_1 \\ n_{21} & a_{22} & a_{23} & a_{24} & a_{25} & b_2 \\ n_{31} & n_{32} & a_{33} & a_{34} & a_{35} & b_3 \\ n_{41} & n_{42} & a_{43} & a_{44} & a_{45} & b_4 \\ n_{51} & n_{52} & a_{53} & a_{54} & a_{55} & b_5 \end{bmatrix}. \quad (34.1)$$

Элементы $1', 2', \dots$ в последней строке дают номера ведущих строк, причем r' определялось на $(r - 1)$ -м шаге. Способ описания очень близок к тому, который используется в большинстве алгоритмических языков и в котором, например, через a_{ij} обозначается соответствующий элемент в позиции (i, j) , полученный после выполнения всех операций предыдущих шагов. Мы опустим верхний индекс, который использовался в математическом описании. Тогда r -й основной шаг состоит в следующем:

(i) Переставляем a_{rj} ($j = r, \dots, n$) и b_r с $a_{r'j}$ ($j = r, \dots, n$) и $b_{r'}$. Заметим, что если $r' = r$, то перестановки нет. Для каждого значения s от $(r + 1)$ до n последовательно выполняем (ii), (iii), (iv).

(ii) Вычисляем $n_{sr} = a_{sr}/a_{rr}$ и записываем на место a_{sr} .

(iii) Для каждого значения t от $(r + 1)$ до n вычисляем $a_{st} - n_{sr}a_{rt}$ и записываем на место a_{st} .

(iv) Вычисляем $b_s - n_{sr}b_r$ и записываем на место b_s . Во время выполнения r -го основного шага запоминаем наибольшую из величин $|a_{s, r+1}|$. Если после завершения шага это будет $a_{(r+1)'}$, $r+1$, то $(r + 1)'$ записывается в $(r + 1)$ -й позиции последней строки.

Очевидно, что предварительно нужно определить максимальный по модулю элемент в первом столбце A_0 .

Решения, соответствующие различным правым частям

35. Нам часто нужно будет решать системы уравнений $Ax = b$ с одной и той же матрицей A , но с различными правыми частями. Если правые части известны с самого начала, то очевидно, что мы можем обработать их все во время процесса исключения. Заметим, что если мы возьмем n правых частей $e_1, \dots, e_n$, то n решений будут столбцами матрицы, обратной к A .

Однако часто мы будем использовать такие методы, в которых каждая правая часть определяется из решения, полученного с предыдущей правой частью. Отдельная обработка правой части состоит из $(n - 1)$ -го основного шага, из которых r -й шаг заключается в следующем:

(i) переставляем b_r и $b_{r'}$;

(ii) для каждого значения s от $(r + 1)$ до n вычисляем $b_s - n_{sr}b_r$ и записываем на место b_s .

Прямое треугольное разложение

36. Вернемся на некоторое время к процессу исключения без перестановок. Очевидно, что промежуточные матрицы A_r сами по себе имеют мало значения, и основная цель процесса — получить матрицы N и A_{n-1} такие, что

$$NA_{n-1} = A_0. \quad (36.1)$$

Если даны N и A_{n-1} , то решение $Ax = b$ с любой правой частью b может быть сведено к решению двух треугольных систем

$$Ny = b, \quad A_{n-1}x = y. \quad (36.2)$$

Это наводит на мысль, что нужно рассмотреть возможность получения матриц N и A_{n-1} непосредственно из A_0 без промежуточных этапов.

Так как A_r должны быть исключены, нет никакого смысла пользоваться нашей старой системой обозначений. Поэтому мы рассмотрим проблему определения нижней треугольной матрицы L с единичными

диагональными элементами, верхней треугольной матрицы U и вектора c таких, что

$$L(U \dot{c}) = (A \dot{b}). \quad (36.3)$$

Предположим, что первые $(r - 1)$ строк L , U и c могут быть определены приравниванием первых $(r - 1)$ строк обеих частей уравнения (36.3). Тогда, приравнявая элементы r -й строки, имеем

$$\left. \begin{aligned} l_{r1}u_{11} &= a_{r1}, \\ l_{r1}u_{12} + l_{r2}u_{22} &= a_{r2}, \\ \dots &\dots \\ l_{r1}u_{1r} + l_{r2}u_{2r} + \dots + l_{rr}u_{rr} &= a_{rr}, \end{aligned} \right\} \quad (36.4)$$

$$\left. \begin{aligned} l_{r1}u_{1, r+1} + l_{r2}u_{2, r+1} + \dots + l_{rr}u_{r, r+1} &= a_{r, r+1}, \\ \dots &\dots \\ l_{r1}u_{1n} + l_{r2}u_{2n} + \dots + l_{rr}u_{rn} &= a_{rn}, \\ l_{r1}c_1 + l_{r2}c_2 + \dots + l_{rr}c_r &= b_r. \end{aligned} \right\} \quad (36.5)$$

Из первых $(r - 1)$ уравнений (36.4) величины l_{r1} , l_{r2} , $\dots$, l_{rr} определяются единственным образом. При произвольном выборе l_{rr} r -е уравнение определяет u_{rr} . Тогда оставшиеся уравнения однозначно определяют величины от $u_{r, r+1}$ до u_{rn} и c_r . Так как для первой строки результат верен, то он верен и в общем случае.

Мы видим, что произведение $l_{rr}u_{rr}$ определяется однозначно, но или l_{rr} , или u_{rr} могут выбираться произвольно. Если мы потребуем, чтобы L была нижней треугольной матрицей с единичными диагональными элементами, то $l_{rr} = 1$. В этом случае деления имеют место только при определении l_{ri} ($i = 1, \dots, r - 1$) и делителями являются u_{ii} . Разложение не может не осуществиться, если $u_{ii} \neq 0$, и тогда оно единственно. Элементы определяются в следующем порядке:

первая строка U , c ; вторая строка L ; вторая строка U , c и т. д. (36.6)
Легко проверить, что они также могут быть определены в таком порядке:

первая строка U , c ; первый столбец L ; вторая строка U , c ; второй столбец L и т. д. (36.7)

Связь метода Гаусса с прямым треугольным разложением

37. Мы уже показали, что если A — невырожденная, то треугольное разложение единственно (при условии $l_{rr} = 1$), если оно существует, и, следовательно, L и U должны совпадать с N и A_{n-1} , полученными в методе Гаусса. Поэтому несостоятельность и неединственность разложения объясняются теми же самыми обстоятельствами, что и в методе Гаусса без выбора главного элемента.

Рассмотрим, например, определение l_{r4} из соотношения

$$l_{r4} = (a_{r4} - l_{r1}u_{14} - l_{r2}u_{24} - l_{r3}u_{34})/u_{44}. \quad (37.1)$$

Сравнение с методом Гаусса показывает, что:

$$\left. \begin{aligned} \text{(i)} \quad a_{r4}^{(0)} &= a_{r4}, \\ \text{(ii)} \quad a_{r4}^{(1)} &= a_{r4} - l_{r1}u_{14}, \\ \text{(iii)} \quad a_{r4}^{(2)} &= a_{r4} - l_{r1}u_{14} - l_{r2}u_{24}, \\ \text{(iv)} \quad a_{r4}^{(3)} &= a_{r4} - l_{r1}u_{14} - l_{r2}u_{24} - l_{r3}u_{34}, \end{aligned} \right\} \quad (37.2)$$

и, следовательно, при определении числителя (37.1) мы получаем по очереди каждый из элементов в позиции $(r, 4)$ для A_1, A_2, A_3 . Аналогичное замечание справедливо и для вычисления элементов U .

Однако если используется треугольное разложение, то мы не записываем эти промежуточные результаты. Более того, если у нас есть возможность для применения fi_2 или fl_2 вычислений, то и не нужно округлять результат после каждого сложения. Числитель может быть накоплен в таком виде и разделен на u_{44} . *Но если мы используем обычную арифметику с плавающей запятой или арифметику с фиксированной запятой без накопления, то ошибки округления в обоих процессах совпадают.*

Примеры несостоятельности и неединственности разложения

38. Пример матрицы, для которой треугольное разложение невозможно, таков:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ 4 & 6 & 7 \end{bmatrix}. \quad (38.1)$$

Элементы L и U , полученные до того шага, когда дальнейшее разложение стало невозможно, такие:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & ? & . \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & -5 \\ . & . & . \end{bmatrix}. \quad (38.2)$$

Мы видим, что $u_{22} = 0$, и уравнение, определяющее l_{32} , есть

$$4 \times 2 + l_{32} \times 0 = 6.$$

Заметим, что A — невырожденная. В действительности она хорошо обусловлена.

С другой стороны, существует бесконечное число разложений матрицы A ,

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 1 \\ 3 & 3 & 1 \end{bmatrix}. \quad (38.3)$$

Например,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 3 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}. \quad (38.4)$$

Опять $u_{22} = 0$, но когда мы приходим к определению l_{32} , то

$$3 \times 1 + l_{32} \times 0 = 3$$

и, следовательно, l_{32} произвольно. Если l_{32} выбрано, то u_{33} определяется однозначно. Эквивалентная треугольная система уравнений, соответствующая произвольной правой части, есть

$$\left. \begin{aligned} x_1 + x_2 + x_3 &= c_1, \\ -x_3 &= c_2, \\ x_3 &= c_3. \end{aligned} \right\} \quad (38.5)$$

Следовательно, если c_2 не равно $-c_3$, то решение не существует. Заметим, что если L должна иметь единичную диагональ, то *необходимым* условием неопределенности является равенство нулю по крайней мере одного u_{rr} . Если тем не менее треугольное разложение существует, то A должна быть вырожденной, так как $A = LU$, а $\det U$ равен нулю. Мы знаем, что решение существует тогда и только тогда, когда

$$\text{ранг}(A \parallel b) = \text{ранг}(A).$$

В том случае, когда единственный нулевой элемент есть u_{nn} , разложение существует, неопределенности нет, но решение возможно лишь при $c_n = 0$.

Треугольное разложение с перестановкой строк

39. Так как исключение Гаусса и треугольное разложение эквивалентны, за исключением, возможно, ошибок округления, то очевидно, что для численной устойчивости разложения необходим выбор главного элемента. На автоматических вычислительных машинах более удобно выполнять эквивалент выбора элемента по столбцу, а не по всей матрице, если мы не хотим использовать вдвое больше памяти или же выполнять почти все операции дважды.

Мы опишем здесь эквивалент выбора главного элемента по столбцу в арифметике с фиксированной запятой. Он состоит из n основных шагов, причем на r -м шаге мы определяем r -й столбец L , r -ю строку $(U \parallel c)$ и r' . Состояние в начале r -го шага иллюстрируется случаем $n = 5$, $r = 3$:

$$\left[\begin{array}{ccccc|cc} u_{11} & u_{12} & u_{13} & u_{14} & u_{15} & c_1 & s_1 \\ l_{21} & u_{22} & u_{23} & u_{24} & u_{25} & c_2 & s_2 \\ l_{31} & l_{32} & a_{33} & a_{34} & a_{35} & b_3 & s_3 \\ l_{41} & l_{42} & a_{43} & a_{44} & a_{45} & b_4 & s_4 \\ l_{51} & l_{52} & a_{53} & a_{54} & a_{55} & b_5 & s_5 \end{array} \right] \cdot \quad (39.1)$$

1' 2'

Теперь r -й основной шаг заключается в следующем.

(i) Для $t = r, r+1, \dots, n$:

Вычисляем $a_{tr} - l_{t1}u_{1r} - l_{t2}u_{2r} - \dots - l_{t,r-1}u_{r-1,r}$, помещая результат в конце t -й строки как число s_t с двойной точностью. (Заметим, что если используется накопление с плавающей запятой, то теряется немного, если s_t округляется после вычисления.)

(ii) Предположим, что $|s_{r'}|$ есть максимум из $|s_t|$ ($t = r, \dots, n$). Тогда записываем r' и переставляем целиком строки r и r' , включая l_{ri} , a_{ri} , b_r , s_r . Округляем новое s_r до одинарной точности, чтобы получить u_{rr} , и записываем его на место a_{rr} .

(iii) Для $t = r+1, \dots, n$:

Для получения l_{tr} вычисляем s_t/u_{rr} и записываем его на место a_{tr} .

(iv) Для $t = r+1, \dots, n$:

Вычисляем $a_{rt} - l_{r1}u_{1t} - l_{r2}u_{2t} - \dots - l_{r,r-1}u_{r-1,t}$, накапливая скалярное произведение и округляя его до одинарной точности, чтобы получить u_{rt} . Записываем u_{rt} на место a_{rt} .

(v) Вычисляем $b_r - l_{r1}c_1 - l_{r2}c_2 - \dots - l_{r,r-1}c_{r-1}$, снова накапливая скалярное произведение и округляя его, чтобы получить c_r . Записываем c_r на место b_r .

Очевидно, что элементы L и U находятся в порядке, указанном в (36.7). Окончательный результат таков, что LU равно A с переставленными строками, и мы сохранили информацию для отдельной обработки правой части. В этой обработке n шагов. На r -м шаге мы переставляем r -ю и r' -ю строки правой части и затем выполняем операцию (v) . На получение каждого из элементов s_r влияет лишь одна ошибка округления.

Заметим, что мы могли бы организовать обработку правой части таким же образом и для исключения Гаусса. Она включала бы перестановку целых строк r и r' на шаге (i) из § 34. Тогда мы могли бы выполнить отдельную обработку правой части с накоплением скалярных произведений.

В табл. 4 мы показываем треугольное разложение с перестановками для матрицы 4-го порядка с переставленными строками из табл. 1.

Таблица 4

*Треугольное разложение с перестановками**Исходные уравнения*

$$\begin{array}{cccc|c} 0,7321 & 0,4135 & 0,3126 & 0,5163 & 0,8132 \\ 0,2317 & 0,6123 & 0,4137 & 0,6696 & 0,4753 \\ 0,4283 & 0,8176 & 0,4257 & 0,8312 & 0,2167 \\ 0,8653 & 0,2165 & 0,8265 & 0,7123 & 0,5165 \end{array}$$

$1' = 1$

Конфигурация в конце первого основного шага

$$0,8653 \quad 0,2165 \quad 0,8265 \quad 0,7123 \quad | \quad 0,5165$$

$$\begin{array}{cccc|c} 0,2678 & 0,6123 & 0,4137 & 0,6696 & 0,4753 \\ 0,4950 & 0,8176 & 0,4257 & 0,8312 & 0,2167 \\ 0,8461 & 0,4135 & 0,3126 & 0,5163 & 0,8132 \end{array}$$

$$s_2 = 0,6123 - (0,2678)(0,2165) = 0,55432130$$

$$s_3 = 0,8176 - (0,4950)(0,2165) = 0,71043250$$

$$s_4 = 0,4135 - (0,8461)(0,2165) = 0,23031935$$

$$2' = 3$$

(4)

Конфигурация в конце второго основного шага

$$\begin{array}{cccc|c} 0,8653 & 0,2165 & 0,8265 & 0,7123 & 0,5165 \\ 0,4950 & 0,7104 & 0,0166 & 0,4786 & -0,0390 \end{array}$$

$$\begin{array}{cccc|c} 0,2678 & 0,7803 & 0,4137 & 0,6696 & 0,4753 \\ 0,8461 & 0,3242 & 0,3126 & 0,5163 & 0,8132 \end{array}$$

$$s_3 = 0,4137 - (0,2678)(0,8265) - (0,7803)(0,0166) = -0,17941032$$

$$s_4 = 0,3126 - (0,8461)(0,8265) - (0,3242)(0,0166) = -0,39208337$$

$$3' = 4$$

(4)

(3)

Конфигурация в конце третьего основного шага

$$\begin{array}{cccc|c} 0,8653 & 0,2165 & 0,8265 & 0,7123 & 0,5165 \\ 0,4950 & 0,7104 & 0,0166 & 0,4786 & -0,0390 \\ 0,8461 & 0,3242 & -0,3921 & -0,2415 & 0,3888 \end{array}$$

$$\begin{array}{cccc|c} 0,2678 & 0,7803 & -0,4576 & 0,6696 & 0,4753 \end{array}$$

(4)

(3)

(4)

$$s_4 = 0,6696 - (0,2678)(0,7123) - (0,7803)(0,4786) - (-0,4576)(-0,2415) = 0,00511592$$

$$4' = 4$$

Продолжение табл. 4

Конфигурация в конце четвертого основного шага				x	r (точное)
0,8653	0,2165	0,8265	0,7123	0,5165	—0,02332
0,4950	0,7104	0,0166	0,4786	—0,0390	0,03339
0,8461	0,3242	—0,3921	—0,2415	0,3888	0,03005
0,2678	0,7803	—0,4576	—0,0051	0,5453	—0,01166
(4)	(3)	(4)	(4)		

 $L \times (U \vdash c)$ (точное). (Сравни с $(A \vdash b)$ с переставленными строками.)

0,86530000	0,21650000	0,82650000	0,71230000	0,51650000
0,42832350	0,81756750	0,42571750	0,83118850	0,21666750
0,73213033	0,41349233	0,31258337	0,51633915	0,81316685
0,23172734	0,61230382	0,41371464	0,66961592	0,47527212

Приводится вид результатов в конце каждого основного шага; элементы матрицы L даны полужирным шрифтом, а элементы матрицы, которые еще не изменялись (за исключением перестановок), окружены пунктирной линией.

Анализ ошибок треугольного разложения

40. Как и в методе Гаусса с выбором главного элемента по столбцу, мы могли бы ожидать, что максимум $|u_{ij}|$, полученных по алгоритму последнего параграфа, редко должен существенно превышать максимум $|a_{ij}|$, и в действительности для плохо обусловленных матриц $|u_{ij}|$ обычно будут меньше, чем $|a_{ij}|$.

Если необходимо, то дополнительный контроль величины элементов может быть достигнут способом, аналогичным способу § 28. Мы нормируем A так, что $|a_{ij}| < 1/2$, и при накоплении s_t и u_{rt} проверяем скалярное произведение после каждого сложения; если по абсолютному значению оно превышает $1/2$, то мы делим s_t или u_{rt} и все элементы A и b в строке t или r на 2. Заметим, что во время накопления s_t или u_{rt} может потребоваться более чем одно такое деление. Как и в методе Гаусса, мы надеемся, что эти деления будут встречаться редко.

Мы знаем, что за исключением ошибок округления $L(U \vdash c)$ равно $(A \vdash b)$ с переставленными строками. Чтобы упростить запись, будем использовать $(A \vdash b)$ для обозначения этой переставленной матрицы, так что теперь нам нужна оценка для $L(U \vdash c) - (A \vdash b)$. Предполагая, что деления на 2 не были необходимыми, имеем

$$l_{tr} \equiv s_t / u_{rr} + \xi_{tr} \quad \left(|\xi_{tr}| \leq \frac{1}{2} 2^{-t} \right), \quad (40.1)$$

где опять l_{tr} , s_t , u_{rr} относятся к вычисленным значениям. Следовательно,

$$a_{tr} \equiv l_{t1}u_{1r} + l_{t2}u_{2r} + \dots + l_{t,r-1}u_{r-1,r} + l_{tr}u_{rr} + u_{rr}\xi_{tr}. \quad (40.2)$$

Аналогично для вычисленных u_{rt} и c_r получаем

$$u_{rt} \equiv a_{rt} - l_{r1}u_{1t} - l_{r2}u_{2t} - \dots - l_{r,r-1}u_{r-1,t} + \varepsilon_{tr} \quad \left(|\varepsilon_{tr}| \leq \frac{1}{2} 2^{-t} \right), \quad (40.3)$$

$$c_r \equiv b_r - l_{r1}c_1 - l_{r2}c_2 - \dots - l_{r,r-1}c_{r-1} + \varepsilon_r \quad \left(|\varepsilon_r| \leq \frac{1}{2} 2^{-t} \right). \quad (40.4)$$

Собирая члены L , U и c в одну сторону, из (40.2), (40.3), (40.4) находим, что

$$L(U \vdash c) \equiv (A + F, b + k), \quad (40.5)$$

$$|f_{ij}| \leq \left\{ \begin{array}{ll} \frac{1}{2} 2^{-t} & (i \leq j), \\ \frac{1}{2} |u_{jj}| 2^{-t} & (i > j), \end{array} \right\} \quad (40.6)$$

$$|k_i| \leq \frac{1}{2} 2^{-t}.$$

Так как мы предполагаем, что $|u_{ij}| \leq 1$, то видим, что все f_{ij} удовлетворяют неравенству $|f_{ij}| \leq \frac{1}{2} 2^{-t}$, но если некоторые $|u_{jj}|$ много меньше единицы, то многие из элементов $|f_{ij}|$ будут существенно меньше, чем $\frac{1}{2} 2^{-t}$.

Соотношения (40.6) показывают, что треугольное разложение с перестановками должно быть существенно более точным процессом, если мы можем накапливать скалярные произведения. В том случае, когда элементы A не представлены точно t -разрядными числами, ошибки, сделанные при первоначальном округлении до t разрядов, обычно влияют так же, как и все ошибки округления в триангуляризации. Заметим, что мы вынуждены сказать «обычно», для того чтобы учесть те редкие случаи, когда даже при использовании перестановок некоторые элементы U существенно больше элементов A .

Внизу табл. 4 дано точное произведение L и $(U \vdash c)$ для сравнения с $(A \vdash b)$. Видно, что максимальная разность равна 0,00003915 и соответствует элементу (1, 4). Она меньше, чем максимальное значение одной ошибки округления.

Вычисление определителей

41. Из соотношения $LU = A + F$, где A отличается от исходной матрицы лишь перестановкой строк, выводим, что

$$\det(A + F) = u_{11}u_{22} \dots u_{nn}. \quad (41.1)$$

Для учета первоначального порядка мы должны умножить на $(-1)^k$, где k есть число значений r , для которых $r' \neq r$. На практике определители почти всегда вычисляются с использованием плавающей арифметики. Поэтому вычисленное значение D таково:

$$D = (1 + \varepsilon) \prod u_{rr} \quad (|\varepsilon| \leq (n-1) 2^{-t}). \quad (41.2)$$

Для любого приемлемого значения n множитель $(1 + \varepsilon)$ соответствует очень малой относительной ошибке. С точностью до этого множителя D есть точное значение $\det(A + F)$. Если λ'_i являются собственными значениями $(A + F)$ и λ_i — собственными значениями A , то

$$\det A = \prod \lambda_i, \quad D = (1 + \varepsilon) \prod \lambda'_i. \quad (41.3)$$

Эти результаты важны в связи с практической проблемой собственных значений.

Разложение Холецкого

42. Если матрица A симметричная и положительно определенная, то диагональные элементы L могут быть выбраны таким образом, чтобы L была вещественной и $U = L^T$. Доказательство проведем по индукции. Предположим, что это верно для матриц до $(n - 1)$ -го порядка. Теперь, если A_n есть положительно определенная матрица порядка n , мы можем написать

$$A_n = \left[\begin{array}{c|c} A_{n-1} & b \\ \hline b^T & a_{nn} \end{array} \right]; \quad (42.1)$$

где A_{n-1} — матрица главного минора порядка $(n - 1)$ и поэтому положительно определенная (гл. 1, §27). Следовательно, по предположению, существует матрица L_{n-1} такая, что

$$L_{n-1}L_{n-1}^T = A_{n-1}. \quad (42.2)$$

Очевидно, что L_{n-1} невырожденная, так как $[\det L_{n-1}]^2 = \det A_{n-1}$, поэтому существует вектор c такой, что

$$L_{n-1}c = b. \quad (42.3)$$

Для любого значения x

$$\left[\begin{array}{c|c} L_{n-1} & 0 \\ \hline c^T & x \end{array} \right] \left[\begin{array}{c|c} L_{n-1}^T & c \\ \hline 0 & x \end{array} \right] = \left[\begin{array}{c|c} A_{n-1} & b \\ \hline b^T & c^T c + x^2 \end{array} \right], \quad (42.4)$$

так что, если мы определим x соотношением

$$c^T c + x^2 = a_{nn}, \quad (42.5)$$

то имеем треугольное разложение A_n . Чтобы показать, что x — вещественное, возьмем определители от обеих частей (42.4), тогда

$$[\det L_{n-1}]^2 x^2 = \det A_n. \quad (42.6)$$

Правая часть положительная, потому что A_n положительно определена, и, следовательно, x — вещественное и может быть взято положительным. Утверждение доказано, так как оно верно для $n = 1$. Заметим, что

$$c_1^2 + c_2^2 + \dots + c_{n-1}^2 + x^2 = a_{nn}. \quad (42.7)$$

Итак, если $|a_{ij}| < 1$, то $|l_{ij}| < 1$, и далее

$$l_{i1}^2 + l_{i2}^2 + \dots + l_{ii}^2 = a_{ii} < 1, \quad (42.8)$$

что показывает, что симметричное разложение может быть выполнено в арифметике с фиксированной запятой. Очевидно, что никаких перестановок не требуется. Это симметричное разложение предложено Холецким.

Произвольные симметричные матрицы

43. Если A — симметричная, но не положительно определенная, то треугольное разложение без перестановок может не существовать. Простой пример иллюстрируется матрицей

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \quad (43.1)$$

Даже если треугольное разложение существует, оно не будет вида LL^T с вещественной невырожденной L , потому что для таких L матрица LL^T обязательно положительно определенная. Однако если треугольные разложения все же существуют, то будут существовать и разложения вида LL^T , если мы допустим использование комплексной арифметики. Интересно отметить, что каждый столбец L или целиком вещественный или целиком чисто мнимый, так что можем написать

$$L_n = M_n D_n, \quad (43.2)$$

где M_n — вещественная, а D_n — диагональная матрица, каждый элемент которой равен 1 или i . Доказательство опять по индукции и почти не отличается от доказательства § 42. Значение x^2 снова вещественное, но может быть отрицательным.

Мы можем избежать использования мнимых величин, если заметим, что

$$A_n = L_n L_n^T = M_n D_n D_n^T M_n^T. \quad (43.3)$$

Здесь матрица D_n^2 диагональная, элементы которой равны ± 1 , и если напомним

$$A_n = (M_n D_n D_n) M_n^T = N_n M_n^T, \quad (43.4)$$

то это дает треугольное разложение A_n , где каждый столбец N_n равен или соответствующей строке M_n^T или той же строке с элементами, знаки которых изменены на обратные. Мы можем связать знаки матрицы D_n с N_n и M_n таким образом, чтобы все диагональные элементы M_n были положительны. Тогда обе матрицы полностью определяются матрицей N_n . Если, например,

$$N_n^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ 1 & 2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \text{то} \quad M_n^T = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & -2 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (43.5)$$

Нельзя с полной уверенностью утверждать, что симметричное разложение неположительно определенной матрицы с точки зрения численной устойчивости не имеет ничего общего со случаем положительно определенной матрицы. Но чтобы быть уверенными в устойчивости, мы должны использовать перестановки, а это нарушает симметрию. Хотя выбор наибольшего диагонального элемента сохраняет симметрию, он не гарантирует устойчивости.

Анализ ошибок разложения Холецкого в арифметике с фиксированной запятой

44. Теперь мы ограничимся случаем положительно определенной матрицы и покажем, что если

$$(i) \quad |a_{ij}| < 1 - (1,00001) 2^{-t} \quad \text{и} \quad (ii) \quad \lambda_n = 1/\|A^{-1}\| > \frac{1}{2} (n+2) 2^{-t}, \quad (44.1)$$

то вычисленное разложение Холецкого в арифметике с фиксированной запятой с накоплением скалярных произведений дает L , удовлетворяю-

щую соотношениям

$$LL^T = A + F, \quad |l_{ij}| \leq 1, \quad (44.2)$$

$$|f_{rs}| \leq \begin{cases} \frac{1}{2} l_{rr} 2^{-t} \leq \frac{1}{2} 2^{-t} & (r > s), \\ \frac{1}{2} l_{ss} 2^{-t} \leq \frac{1}{2} 2^{-t} & (r < s), \\ (1,00001) l_{rr} 2^{-t} \leq (1,00001) 2^{-t} & (r = s). \end{cases} \quad (44.3)$$

Доказательство по индукции. Предположим, что мы вычислили элементы L до $l_{r, s-1}$ ($s-1 < r$), и эти элементы соответствуют элементам точного разложения $(A + F_{r, s-1})$, где $F_{r, s-1}$ — симметричная и в позициях

$$i, j \leq r-1; \quad i=r, \quad j \leq s-1; \quad j=r, \quad i \leq s-1$$

имеет элементы, удовлетворяющие соотношениям (44.2) и (44.3); в других местах матрица $F_{r, s-1}$ нулевая. Тогда имеем

$$\begin{aligned} \lambda_{\min}(A + F_{r, s-1}) &> \lambda_{\min}(A) + \lambda_{\min}(F_{r, s-1}) > \\ &> \lambda_n - \frac{1}{2} (r+2) 2^{-t} > \\ &> 0 \quad (\text{см. (ii)}), \end{aligned} \quad (44.4)$$

и $|(A + F_{r, s-1})_{ij}| \leq 1$. Следовательно, $(A + F_{r, s-1})$ — положительно определенная и имеет элементы, ограниченные по модулю единицей. Поэтому она имеет точное треугольное разложение, у которого мы уже получили элементы до $l_{r, s-1}$. Теперь рассмотрим выражение

$$\frac{a_{rs} - l_{r1}l_{s1} - l_{r2}l_{s2} - \dots - l_{r, s-1}l_{s, s-1}}{l_{ss}} = x, \quad (44.5)$$

где l_{ij} — вычисленные элементы. Это есть элемент (r, s) треугольного разложения матрицы $(A + F_{r, s-1})$ и, следовательно, он ограничен по модулю единицей. По определению $fi_2(x)$ есть правильно округленное значение x и поэтому не может превзойти по модулю единицу. Тогда для вычисленного l_{rs} имеем

$$l_{rs} = \left(\frac{a_{rs} - l_{r1}l_{s1} - l_{r2}l_{s2} - \dots - l_{r, s-1}l_{s, s-1}}{l_{ss}} \right) + \xi_{rs} \quad \left(|\xi_{rs}| \leq \frac{1}{2} 2^{-t} \right). \quad (44.6)$$

Это означает, что

$$(LL^T)_{rs} = a_{rs} + l_{ss}\xi_{rs} \equiv a_{rs} + f_{rs}, \quad (44.7)$$

где

$$|f_{rs}| \leq \frac{1}{2} l_{ss} 2^{-t} \leq \frac{1}{2} 2^{-t}, \quad (44.8)$$

и мы установили нужные неравенства для элементов до l_{rs} . Чтобы закончить индуктивное доказательство, нужно рассмотреть вычисление диагонального элемента l_{rr} . Если

$$y^2 = a_{rr} - l_{r1}^2 - l_{r2}^2 - \dots - l_{r, r-1}^2, \quad (44.9)$$

то y есть (r, r) -й элемент точного треугольного разложения $(A + F_{r, r-1})$ и поэтому лежит в пределах от 0 до 1. Следовательно,

$$l_{rr} = fi_2[(a_{rr} - l_{r1}^2 - l_{r2}^2 - \dots - l_{r, r-1}^2)^{1/2}] \equiv \\ \equiv y + \varepsilon \quad (|\varepsilon| < 1,00001 \cdot 2^{-t-1}; \text{ см. гл. 3, § 10}), \quad (44.10)$$

где мы предполагаем, что программа вычисления квадратного корня всегда дает значение, не большее единицы. Следовательно,

$$l_{rr}^2 \equiv y^2 + 2\varepsilon y + \varepsilon^2$$

или

$$l_{r1}^2 + l_{r2}^2 + \dots + l_{rr}^2 \equiv a_{rr} + 2\varepsilon y + \varepsilon^2 \leq a_{rr} + 2\varepsilon y + 2\varepsilon^2 = a_{rr} + 2\varepsilon l_{rr}, \quad (44.11)$$

откуда

$$(LL^T)_{rr} \equiv a_{rr} + f_{rr} \quad (|f_{rr}| \leq (1,00001) l_{rr} 2^{-t}). \quad (44.12)$$

Итак, результат полностью установлен.

Аналогичный результат может быть установлен и в том случае, если используется fl_2 ().

Пренебрегая членами порядка 2^{-2t} , можно получить такие оценки:

$$|f_{rs}| \leq \begin{cases} |l_{rs} l_{ss}| 2^{-t} & (r > s), \\ |l_{rs} l_{rr}| 2^{-t} & (r < s), \\ l_{rr}^2 2^{-t} & (r = s). \end{cases} \quad (44.13)$$

Плохо обусловленная матрица

45. Если большинство из элементов l_{rs} существенно меньше единицы, то эти оценки означают, что LL^T отличается от A в большинстве своих элементов на величину, существенно меньшую чем $\frac{1}{2} 2^{-t}$, и невозможно, что ошибки, сделанные при разложении, играют меньшую роль, чем ошибки, сделанные при первоначальном округлении элементов матрицы до t -разрядного представления.

В табл. 5 мы показываем треугольное разложение очень плохо обусловленной матрицы H , полученной из подматрицы пятого порядка матрицы Гильберта, умноженной на 1,8144 для того, чтобы избежать ошибок округления в первоначальном представлении. Вычисления выполнялись с использованием fl_2 () и восьми десятичных разрядов. Приводится разность между LL^T и H ; заметим, что большинство ее элементов по величине много меньше чем $\frac{1}{2} 10^{-8}$. Интересная особенность этого примера

заключается в том, что наименьшие элементы $(LL^T - H)$ находятся в тех позициях, в которых возмущение сильнее всего влияет на решение. Эти результаты можно сравнить с результатами, полученными Уилкинсоном (1961b) для арифметики с фиксированной запятой и накоплением.

Видно, что элементы $(LL^T - H)$ для плавающей запятой меньше, чем для фиксированной. (Исследование обусловленности матрицы Гильберта и некоторых родственных матриц см. у Тодда (1950, 1961).)

Таблица 5

Верхний треугольник Π (симметричный)

$$\begin{bmatrix} 0,90720 & 0,60480 & 0,45360 & 0,36288 & 0,30240 \\ 0,45360 & 0,36288 & 0,30240 & 0,25920 & 0,25920 \\ 0,30240 & 0,25920 & 0,22680 & 0,20160 & 0,18144 \end{bmatrix}$$

 L^T

$$\begin{bmatrix} 10^0 (0,95247047) & 10^0 (0,63498032) & 10^0 (0,47623524) & 10^0 (0,38098819) & 10^0 (0,31749016) \\ 10^0 (0,22449943) & 10^0 (0,26939933) & 10^0 (0,26939933) & 10^0 (0,26939933) & 10^0 (0,25637079) \\ 10^{-1} (0,95247047) & 10^{-1} (0,54990883) & 10^{-1} (0,94270103) & 10^{-1} (0,94270103) & 10^0 (0,41783769) \\ & & 10^{-1} (0,13606703) & 10^{-1} (0,30237004) & 10^{-1} (0,33808914) \\ & & & 10^{-2} (0,33808914) & \end{bmatrix}$$

 $10^8 (LL^T - H)$

$$\begin{bmatrix} 10^0 (0,38311504) & 10^0 (0,28733628) & 10^{-1} (0,3937493) & 10^0 (0,19155752) \\ 10^{-1} (0,8576273) & 10^0 (0,41178587) & 10^0 (0,4074970) & 10^{-1} (-0,4966991) \\ 10^{-2} (0,35426189) & 10^{-2} (0,4728749) & 10^{-3} (0,4728749) & 10^{-2} (0,9258937) \\ & & 10^{-3} (-0,1927482) & 10^{-3} (-0,2891148) \\ & & & 10^{-4} (0,44040996) \end{bmatrix}$$

 $(L^T)^{-1}$

$$\begin{bmatrix} 10^1 (-0,29695696) & 10^1 (0,54554509) & 10^1 (-0,83992292) & 10^2 (0,11736933) \\ 10^1 (0,4453543) & 10^2 (-0,21821801) & 10^2 (0,62994193) & 10^3 (-0,14084519) \\ & 10^2 (0,1818333) & 10^3 (-0,42598835) & 10^3 (0,49296216) \\ & & 10^2 (0,73493189) & 10^3 (-0,65728637) \\ & & & 10^3 (0,29577998) \end{bmatrix}$$

Триангуляризация, использующая матрицы отражения

46. Теперь обратимся к другим методам триангуляризации, существование которых указывалось в § 15. Сначала опишем метод, принадлежащий Хаусхолдеру и основанный на использовании матриц отражения, и чтобы иметь возможность сравнить его с нашим общим анализом, остановимся на $(r+1)$ -м преобразовании. Всего в методе $(n-1)$ основных шагов, и после окончания r -го шага матрица A_r в своих первых r столбцах будет уже верхней треугольной. Мы можем написать

$$A_r = \left[\begin{array}{c|c} U_r & V_r \\ \hline 0 & W_{n-r} \end{array} \right]^r, \quad (46.1)$$

где U_r — верхняя треугольная матрица порядка r . На $(r+1)$ -м шаге используется матрица P_r вида

$$P_r = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2vv^T \end{array} \right]^r, \quad \|v\|_2 = 1, \quad (46.2)$$

вместо матрицы N_{r+1} , использованной в методе Гаусса.

Имеем

$$A_{r+1} = P_r A_r = \left[\begin{array}{c|c} U_r & V_r \\ \hline 0 & (I - 2vv^T) W_{n-r} \end{array} \right], \quad (46.3)$$

так что здесь изменяется лишь W_{n-r} . Мы должны выбрать v так, чтобы первый столбец матрицы $(I - 2vv^T) W_{n-r}$ был нулевым, кроме своего первого элемента. Эту задачу мы рассматривали в гл. 3, §§ 37–45. Если обозначим элементы первого столбца W_{n-r} через

$$a_{r+1, r+1}, a_{r+2, r+1}, \dots, a_{n, r+1}, \quad (46.4)$$

опуская для простоты верхний индекс, то можем написать

$$I - 2vv^T = I - uu^T / 2K^2, \quad (46.5)$$

где

$$\left. \begin{aligned} u_1 &= a_{r+1, r+1} \pm S, & u_i &= a_{r+i, r+1} & (i = 2, 3, \dots, n-r), \\ S^2 &= \sum_{i=1}^{n-r} a_{r+i, r+1}^2, & 2K^2 &= S^2 \mp a_{r+1, r+1} S = \mp u_1 S. \end{aligned} \right\} \quad (46.6)$$

В этом случае новый элемент $(r+1, r+1)$ равен $\pm S$. Численная устойчивость достигается выбором знака так, чтобы

$$|u_1| = |a_{r+1, r+1}| + S. \quad (46.7)$$

Матрица $(I - 2vv^T) W_{n-r}$ вычисляется из выражения

$$(I - uu^T / 2K^2) W_{n-r}$$

по шагам:

$$p^T = u^T W_{n-r} / 2K^2, \quad (46.8)$$

$$(I - 2vv^T) W_{n-r} = W_{n-r} - up^T. \quad (46.9)$$

Если во время триангуляризации нам известна правая часть, то на $(r+1)$ -м основном шаге мы можем умножить ее на P_r . Однако если мы

хотим решать уравнение с правой частью, которая определяется после триангуляризации, то должны сохранить всю необходимую информацию. Наиболее удобно хранить $(n - r)$ элементов вектора u на месте, занимаемом первым столбцом W_{n-r} . Так как из (46.6) $u_i = a_{r+i, r+1}$ ($i = 2, \dots, n - r$), то все элементы, кроме u_1 , уже на месте. Это означает, что мы дополнительно должны запомнить диагональные элементы верхней треугольной матрицы; они несколько иначе, чем остальные элементы, используются в обратной подстановке, поэтому это не так существенно. Нам также нужны значения, равные $2K^2$, но они могут быть определены из соотношений $2K^2 = \mp u_1 S$. Если принимается последний вариант, то, помимо места для матрицы A_0 , нужно дополнительно иметь n ячеек памяти. Будем называть этот метод *триангуляризацией Хаусхолдера*.

Анализ ошибок триангуляризации Хаусхолдера

47. Общий анализ, который мы дали в гл. 3, §§ 42—45, более чем достаточен, чтобы охватить триангуляризацию Хаусхолдера, но в действительности соответствующие оценки могут быть несколько уменьшены. В § 45 гл. 3 мы получили оценку для $\|\bar{A}_{n-1} - P_{n-2}P_{n-3} \dots P_0 A_0\|_E$, где $\bar{A}_{n-1}$ есть $(n - 1)$ -е *вычисленное* преобразование, а P_{r-1} — точная матрица отражения, соответствующая вычисленной $\bar{A}_r$. (Мы используем здесь черточки, чтобы иметь возможность сравнить с прежним анализом.) Для нашей настоящей цели более уместна оценка для $\bar{A}_{n-1} - \bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0 A_0$. Если

$$\bar{A}_{n-1} - \bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0 A_0 = F, \quad (47.1)$$

то

$$\bar{A}_{n-1} = \bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0 [A_0 + (\bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0)^{-1} F]. \quad (47.2)$$

Отсюда мы видим, что близость $\bar{P}_r$ к матрице отражения нужна лишь для того, чтобы гарантировать, что P_r^{-1} почти ортогональна, и, следовательно, что $\|(\bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0)^{-1} F\|_E$ незначительно больше, чем $\|F\|_E$. Член (9,04) $2^{-t} \|A_0\|_E$ из (42.3) главы 3 не входит в наш настоящий анализ и в соотношении (44.4) главы 3 остается член (3,35) $2^{-t} \|A_0\|_E$. Читателю следует проверить, что

$$\bar{A}_{n-1} = \bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0 (A_0 + G); \quad (47.3)$$

где

$$\|G\|_E \leq (3,35) (n - 1) [1 + (9,01) 2^{-t}]^{n-2} 2^{-t} \|A_0\|_E. \quad (47.4)$$

Аналогично

$$\bar{b}_{n-1} = \bar{P}_{n-2}\bar{P}_{n-3} \dots \bar{P}_0 (b_0 + k); \quad (47.5)$$

$$\|k\|_2 = (3,35) (n - 1) [1 + (9,01) 2^{-t}]^{n-1} 2^{-t} \|b_0\|_2 \quad (47.6)$$

Триангуляризация элементарными устойчивыми матрицами типа M'_{ji}

48. Мы можем заменить одну элементарную устойчивую матрицу N'_r , использованную на r -м основном шаге метода Гаусса с выбором главного элемента по столбцу, на произведение более простых устойчивых матриц

$$M'_{r+1, r}, M'_{r+2, r}, \dots, M'_{n, r}.$$

В этом случае r -й основной шаг распадается на $(n - r)$ элементарных шагов, в каждом из которых получается один нуль. Структура элементов для случая $n = 6$, когда уже выполнены два элементарных шага из третьего основного, такова:

$$\left[\begin{array}{cccccc|c} u_{11} & u_{12} & u_{13} & u_{14} & u_{15} & u_{16} & c_1 \\ * & & & & & & \\ n_{21} & u_{22} & u_{23} & u_{24} & u_{25} & u_{26} & c_2 \\ n_{31} & n_{32} & a_{33} & a_{34} & a_{35} & a_{36} & b_3 \\ * & & & & & & \\ n_{41} & n_{42} & n_{43} & a_{44} & a_{45} & a_{46} & b_4 \\ * & & & & & & \\ n_{51} & n_{52} & n_{53} & a_{54} & a_{55} & a_{56} & b_5 \\ * & & & & & & \\ n_{61} & n_{62} & a_{63} & a_{64} & a_{65} & a_{66} & b_6 \end{array} \right]. \quad (48.1)$$

Мы обозначили элементы в первых двух строках через u_{ij} и c_i , так как это уже окончательные элементы. Значение звездочек станет ясным, когда мы определим r -й шаг следующим образом.

Для каждого значения i от $(r + 1)$ до n :

(i) Сравниваем a_{rr} и a_{ir} . Если $|a_{ir}| > |a_{rr}|$, переставляем элементы a_{rj} и a_{ij} ($j = r, \dots, n$) и b_r и b_i .

(ii) Вычисляем a_{ir}/a_{rr} и записываем как n_{ir} на место a_{ir} . Если на шаге (i) имела место перестановка, то у n_{ir} ставится звездочка.

(iii) Для каждого значения j от $(r + 1)$ до n :

Вычисляем $a_{ij} - n_{ir}a_{rj}$ и записываем на место a_{ij} .

(iv) Вычисляем $b_i - n_{ir}b_r$ и записываем на место b_i .

На практике мы можем пожертвовать последним разрядом в каждом n_{ir} и запоминать «единицу», если имела место перестановка перед вычислением n_{ir} , и «нуль» в противном случае. Очевидно, что мы запомнили достаточное количество информации для раздельной обработки правой части.

Этот алгоритм по устойчивости сравним с методом Гаусса с выбором главного элемента по столбцу, хотя он и не может обеспечить столь малые верхние оценки для эквивалентных возмущений в A_0 . Он непригоден для накопления скалярных произведений и, как мы только что отметили, не дает почти ничего нового по сравнению с методом Гаусса с выбором главного элемента по столбцу.

Вычисление главных миноров

49. Однако шаги алгоритма § 48 могут быть модифицированы таким образом, что он будет иметь существенные преимущества. Модифицированный процесс также имеет $(n - 1)$ основных шагов, но на первых $(r - 1)$ шагах участвуют лишь первые r строк ($A_0 : b_0$). Структура элементов после $(r - 1)$ основных шагов для случая $n = 6$, $r = 3$ такова:

$$\left[\begin{array}{cccccc|c} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & b_1 \\ n_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & b_2 \\ n_{31} & n_{32} & a_{33} & a_{34} & a_{35} & a_{36} & b_3 \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & b_4 \end{array} \right]. \quad (49.1)$$

Строка $(r + 1)$ является неизменной $(r + 1)$ -й строкой A_0 , строки с 1-й до r -й преобразуются способом, который станет ясным из следующего описания r -го основного шага.

Для каждого значения i от 1 до r :

(i) Сравниваем a_{ii} и $a_{r+1, i}$. Если $|a_{r+1, i}| > |a_{ii}|$, переставляем $a_{r+1, j}$ и a_{ij} ($j = i, \dots, n$), и b_{r+1} и b_i .

(ii) Вычисляем $a_{r+1, i}/a_{ii}$ и записываем на место $a_{r+1, i}$ как $n_{r+1, i}$. Если на шаге (i) имела место перестановка, то $n_{r+1, i}$ помечаем звездочкой.

(iii) Для каждого значения j от $(i+1)$ до n :

Вычисляем $a_{r+1, j} - n_{r+1, i}a_{ij}$ и записываем на место $a_{r+1, j}$.

(iv) Вычисляем $b_{r+1} - n_{r+1, i}b_i$ и записываем на место b_{r+1} .

Эта схема имеет два преимущества.

(I) Если мы не хотим сохранить n_{ij} , т. е. если ни одна из правых частей не будет обрабатываться в дальнейшем, то максимум рабочих ячеек памяти, необходимых для реализации данного процесса, например с одной правой частью, равен

$$[(n+1) + n + (n-1) + \dots + 3] + n + 1.$$

(II) С нашей точки зрения более важно то, что мы теперь можем вычислить главный минор A_0 порядка $(r+1)$ на r -м основном шаге и можем остановиться на этом шаге, если того желаем, не затрагивая строк от $(r+2)$ -й до n -й. $(r+1)$ -й главный минор p_{r+1} определяется следующим образом.

Если мы запоминаем текущее число перестановок k , которые имели место с начала первого основного шага, то после выполнения r -го шага имеем:

$$p_{r+1} = (-1)^k a_{11}a_{22} \dots a_{r+1, r+1}, \quad (49.2)$$

где a_{ii} — соответствующие значения.

Часто нам будет нужен лишь знак p_{r+1} . Он может быть получен следующим образом на r -м основном шаге.

Предположим, что мы знаем знак p_r . Тогда каждый раз, когда выполняется описанный выше шаг (i), мы изменяем этот знак, если требовалась перестановка и a_{ii} и $a_{r+1, i}$ имели один и тот же знак. После выполнения r -го основного шага мы снова изменяем знак, если $a_{r+1, r+1}$ отрицательно.

Миноры могут быть вычислены и по алгоритму § 48, но метод, который мы описали, более удобен. Вообще говоря, главные миноры не могут быть вычислены, если мы используем метод Гаусса с выбором главного элемента по столбцу или по всей матрице, или если мы используем триангуляризацию Хаусхолдера.

Триангуляризация плоскими вращениями

50. Наконец, рассмотрим триангуляризацию плоскими вращениями (Гивенс, 1959). Она очень близка к алгоритму, основанному на использовании матриц M_{ji} . Аналогично алгоритму § 48 мы имеем алгоритм, в котором r -й основной шаг состоит из левых умножений на матрицы вращения соответственно в плоскостях $(r, r+1)$, $(r, r+2)$, $\dots$, (r, n) , причем нули в r -м столбце получаются один за одним; r -й основной шаг состоит в следующем (отметим аналогию с § 48).

Для каждого значения i от $(r+1)$ до n :

(i) Вычисляем $x = (a_{rr}^2 + a_{ir}^2)^{1/2}$.

(ii) Вычисляем $\cos \theta = a_{rr}/x$, $\sin \theta = a_{ir}/x$. Если $x = 0$, берем $\cos \theta = 1$, $\sin \theta = 0$. Записываем x на место a_{rr} .

(iii) Для каждого значения j от $(r+1)$ до n :

Вычисляем $a_{rj} \cos \theta + a_{ij} \sin \theta$ и $-a_{rj} \sin \theta + a_{ij} \cos \theta$ и записываем их соответственно на места a_{rj} и a_{ij} .

(iv) Вычисляем $b_r \cos \theta + b_i \sin \theta$ и $-b_r \sin \theta + b_i \cos \theta$ и записываем их соответственно на места b_r и b_i .

Заметим, что для хранения информации о вращении у нас теперь не хватает места в позиции того элемента, который становится нулем, так как нужно разместить как $\cos \theta$, так и $\sin \theta$. Конечно, мы могли бы сохранить лишь один из них, так как другой можно получить, но это нецелесообразно, и если должна быть гарантирована численная устойчивость, нам нужно хранить меньший из них и помнить, который из двух запоминается!

51. Преобразования можно модифицировать таким же способом, как было описано в § 49. В этой модификации строки с $(r+1)$ -й до n -й не изменяются на первых $(r-1)$ шагах. Теперь r -й основной шаг таков.

Для каждого значения i от 1 до r :

(i) Вычисляем $x^2 = (a_{ii}^2 + a_{r+1,i}^2)$.

(ii) Вычисляем $\cos \theta = a_{ii}/x$, $\sin \theta = a_{r+1,i}/x$. Записываем x на место a_{ii} .

(iii) Для каждого значения j от $(i+1)$ до n :

Вычисляем $a_{ij} \cos \theta + a_{r+1,j} \sin \theta$ и $-a_{ij} \sin \theta + a_{r+1,j} \cos \theta$ и записываем их соответственно на места a_{ij} и $a_{r+1,j}$.

(iv) Вычисляем $b_i \cos \theta + b_{r+1} \sin \theta$ и $-b_i \sin \theta + b_{r+1} \cos \theta$ и записываем их соответственно на места b_i и b_{r+1} .

Так как каждая из матриц вращения имеет определитель, равный $+1$, то главный минор $p_{r+1}(r+1)$ -го порядка определяется после выполнения r -го основного шага

$$p_{r+1} = a_{11}a_{22} \dots a_{r+1, r+1}. \quad (51.1)$$

Будем называть триангуляризацию с помощью плоских вращений *триангуляризацией Гивенса*, независимо от того, организована ли она, как в § 50 или так, как мы только что описали.

Анализ ошибок преобразования Гивенса

52. Снова анализ ошибок, который мы дали в гл. 3, §§ 20–36, более чем достаточен, чтобы охватить триангуляризацию Гивенса, и, как в триангуляризации Хаусхолдера, оценки могут быть несколько уменьшены.

Для стандартных вычислений с плавающей запятой мы можем показать, что если окончательно вычисленная система уравнений есть $\bar{A}_N x = \bar{b}_N$, то

$$\bar{A}_N = \bar{R}_N \dots \bar{R}_2 \bar{R}_1 (A_0 + G), \quad (52.1)$$

$$\bar{b}_N = \bar{R}_N \dots \bar{R}_2 \bar{R}_1 (b_0 + g), \quad (52.2)$$

где

$$\|G\|_E \leq \alpha \|A_0\|_E, \quad \|g\|_2 \leq \alpha \|b_0\|_2 \quad (52.3)$$

и

$$\alpha \leq 3n^{3/2} 2^{-t} \left[\frac{1 + 6 \cdot 2^{-t}}{1 - (4,3) 2^{-t}} \right]^{2n-4}. \quad (52.4)$$

Последний множитель в α не имеет особого значения для тех n , для которых α существенно меньше единицы и евклидова норма эквивалентного возмущения в действительности ограничена величиной $kn^{3/2} 2^{-t} \|A_0\|_E$.

Результаты вычислений с фиксированной запятой из главы 3 могут быть несколько улучшены за счет того, что нули, полученные на последовательных шагах, в дальнейшем сохраняются. Далее, простым следствием анализа главы 3 является то, что если каждый столбец первоначально нормирован так, чтобы его 2-норма была меньше чем $1 - \frac{5}{2} n^2 2^{-t}$, то ни один элемент не может дать переполнения во время процесса. Следовательно, если даже мы допустим нормирование лишь с помощью степеней 2, то можем гарантировать, что

$$\frac{1}{2} n^{1/2} \left(1 - \frac{5}{2} n^2 2^{-t} \right) \leq \|A_0\|_E \leq n^{1/2} \left(1 - \frac{5}{2} n^2 2^{-t} \right), \quad (52.5)$$

и при этом не будет опасности переполнения. В действительности с таким нормированием 2-норма каждого из столбцов будет оставаться на каждом шаге меньше единицы. Оценки для эквивалентного возмущения в A_0 могут быть выражены многими способами, но с данным нормированием мы, конечно, имеем

$$\|G\|_E \leq \frac{3}{2} n^{5/2} 2^{-t} < 3n^2 2^{-t} \|A_0\|_E / \left(1 - \frac{5}{2} n^2 2^{-t} \right). \quad (52.6)$$

Аналогично, если так же нормируется и b_0 , то

$$\|g\|_2 < 0,4n^2 2^{-t} < 0,8n^2 2^{-t} \|b_0\|_2. \quad (52.7)$$

Единственность ортогональной триангуляризации

53. Если через Q^T обозначим произведение ортогональных матриц преобразования, полученных из триангуляризации Гивенса, или Хаусхолдера, то

$$Q^T A_0 = U, \quad (53.1)$$

где Q^T — ортогональная, а U — верхняя треугольная. Мы можем записать это в виде

$$A_0 = QU. \quad (53.2)$$

Теперь мы покажем, что если A_0 — невырожденная и представлена в виде (53.2), то Q и U — по существу единственные. Если

$$Q_1 U_1 = Q_2 U_2, \quad (53.3)$$

то

$$Q_2^T Q_1 = U_2 U_1^{-1} = V, \quad (53.4)$$

так что V — также верхняя треугольная. Невырожденность U_1 следует из невырожденности A_0 . Здесь $Q_2^T Q_1$ ортогональная и, следовательно,

$$V^T V = I. \quad (53.5)$$

Приравнивая элементы обеих частей, мы находим, что V должна иметь вид

$$V = D = \text{diag}(d_{ii}), \quad d_{ii} = \pm 1. \quad (53.6)$$

Итак, с точностью до знаков столбцов Q и строк U , разложение (53.2) единственное. Аналогично, если A_0 — комплексная и невырожденная,

(54.3), и матрица, которая могла бы быть получена перемножением транспонированных матриц преобразований процесса Гивенса или Хаусхолдера, могут почти полностью отличаться друг от друга. Вообще говоря, обе последние матрицы могут также отличаться полностью. Из анализа ошибок главы 3 мы знаем, что эти матрицы очень близки к ортогональным для любого приемлемого значения n , хотя на практике их полный вид нам нужен редко. Однако векторы q_i , полученные из применения формул (54.1)–(54.3), могут отклоняться от ортогональных на почти произвольные величины.

На первый взгляд это кажется неожиданным, так как процесс Шмидта вроде бы и построен специально для того, чтобы получать ортогональные векторы. Однако рассмотрим такую матрицу 2-го порядка:

$$A = \begin{bmatrix} 0,63257 & 0,56154 \\ 0,31256 & 0,27740 \end{bmatrix}. \quad (55.1)$$

Первый шаг в процессе Шмидта дает

$$u_{11} = 0,70558, \quad q_1^T = (0,89652, 0,44298), \quad (55.2)$$

второй шаг дает $u_{12} = 0,62631$. Поэтому мы имеем

$$u_{22}q_2^T = a_2^T - u_{12}q_1^T = (0,00004, -0,00004). \quad (55.3)$$

При вычислении правой части (55.3) имеет место значительное сокращение, и полученные компоненты имеют большую относительную ошибку. (Это верно независимо от того, используется ли фиксированная или плавающая запятая.) Следовательно, когда вектор справа нормируется, чтобы получить q_2 , он даже приближенно неортогонален q_1 ! Это вполне убедительно показывает, что данное явление не связано с большим порядком матриц и не может быть отнесено к «совокупному влиянию большого числа ошибок округления».

Естественно спросить, действительно ли «близка» вычисленная ортогональная матрица, полученная по процессу Гивенса или по процессу Хаусхолдера, к той единственной ортогональной матрице, которая соответствует точному вычислению во всех трех методах. В общем случае мы не можем гарантировать, что это действительно так. Наш анализ ошибок из главы 3 гарантирует то, что матрицы будут «почти ортогональны», а не то, что они будут «близки» к матрице, соответствующей точному вычислению. (В гл. 5, § 28 дадим иллюстрацию сказанного в связи с ортогональными подобными преобразованиями.) В некоторых вычислениях ортогональность Q очень важна. Например, в QR -алгоритме для симметричных матриц (гл. 8) после получения разложения матрицы A в произведение QU вычисляется матрица UQ . Так как мы имеем $UQ = Q^{-1}(QU)Q = Q^{-1}AQ$, то эта полученная матрица есть подобное преобразование A , но она будет симметричной только тогда, когда Q ортогональная. Явление, которое мы только что рассмотрели в связи с процессом Шмидта, будет снова интересоваться нас, когда мы рассмотрим метод Арнольди для решения проблемы собственных значений (гл. 6, § 32).

Число умножений, необходимое для выполнения процесса Шмидта, равно n^3 , тогда как триангуляризации Хаусхолдера и Гивенса требуют соответственно $\frac{2}{3}n^3$ и $\frac{4}{3}n^3$. Конечно, процессы Хаусхолдера и Гивенса дают ортогональную матрицу лишь в факторизованной форме, но часто на практике это более удобно. Вычисление произведения преобразований

Хаусхолдера требует дополнительно $\frac{2}{3} n^3$ умножений, так что, даже если ортогональная матрица требуется в явном виде, метод Хаусхолдера вполне экономичен. Более того, если мы хотим получить матрицу, «близкую к ортогональной», используя процесс Шмидта, то векторы должны переортогонализироваться на каждом этапе, а это требует дополнительно n^3 умножений (см. гл. 6).

Сравнение методов триангуляризации

56. Теперь мы впервые подошли к сравнению методов, основанных на устойчивых элементарных преобразованиях и элементарных ортогональных преобразованиях.

Есть один случай, для которого выбор вполне определен — это триангуляризация положительно определенной симметричной матрицы. Здесь симметричное разложение Холецкого обладает всеми достоинствами. Не требуется никаких перестановок, и работа может быть удобно выполнена с фиксированной запятой. Если возможно накопление скалярных произведений, то оценки ошибок при заданной точности вычислений почти не превышают оценок любого другого метода. Он полностью использует симметрию и требует лишь $n^3/6$ умножений. Отметим, что, не учитывая ошибок округления, мы имеем

$$A = LL^T \quad \text{и} \quad A^{-1} = (L^T)^{-1} L^{-1}, \quad (56.1)$$

откуда

$$\|A\|_2 = \|L\|_2^2 \quad \text{и} \quad \|A^{-1}\|_2 = \|L^{-1}\|_2^2, \quad (56.2)$$

и опять здесь самое удовлетворительное состояние дела, так как спектральное число обусловленности матрицы L равно квадратному корню из спектрального числа обусловленности матрицы A . Если A — ленточная матрица, т. е. если $a_{ij} = 0$ для $|i - j| > k$, то $l_{ij} = 0$ для $i - j > k$, так что данная форма используется полностью. Возможно, что единственное замечание, которое мы можем сделать не в пользу метода, состоит в том, что требуется $(n - 1)$ операций извлечения квадратного корня, тогда как в обычном методе Гаусса не требуется ни одной.

57. В более общем случае решение этого вопроса не так ярко выражено. Если матрица совершенно произвольная, то объем работы в четырех методах, которые мы обсудили, приблизительно следующий:

(I) Метод Гаусса с выбором главного элемента по столбцу или по всей матрице: $\frac{1}{3} n^3$ умножений.

(II) Триангуляризация с матрицами типа M'_{ji} : $\frac{1}{3} n^3$ умножений.

(III) Триангуляризация Хаусхолдера: $\frac{2}{3} n^3$ умножений, n квадратных корней.

(IV) Триангуляризация Гивенса: $\frac{4}{3} n^3$ умножений, $\frac{1}{2} n^2$ квадратных корней.

В отношении скорости методы (I) и (II) лучше, чем методы (III) и (IV), и, кроме этого, они несколько проще. На многих вычислительных машинах методы (I) и (II) могут выполняться с двойной точностью почти за то же

самое время, как метод (IV) с одинарной точностью. К сожалению, в отношении численной устойчивости методы (I) и (II) с теоретической точки зрения менее удовлетворительны. Мы видели, что если используем выбор главного элемента по столбцу, то последний ведущий элемент *может* быть в 2^{n-1} раз больше, чем максимальный элемент в исходной матрице и, следовательно, строгие априорные верхние оценки, которые мы можем получить для эквивалентного возмущения исходной матрицы, содержат этот множитель 2^{n-1} . Однако на основе опыта можно заключить, что хотя такая оценка и достижима, она совсем нереальна для практических случаев. В действительности какое-либо возрастание вообще редко имеет место, и если может накапливаться скалярное произведение, то эквивалентные возмущения в исходной матрице исключительно малы.

Для метода Гаусса с выбором главного элемента по всей матрице у нас есть значительно меньшие оценки для максимального роста главного элемента и большие основания думать, что эти оценки не достигаются. К сожалению, по-видимому, невозможно использовать столь выгодное для этого метода накопление скалярных произведений и, следовательно, на вычислительных машинах, имеющих такую возможность, выбор главного элемента по столбцу предпочтительнее, несмотря на потенциальную опасность. К счастью, для некоторых типов матриц, которые имеют большое значение в проблеме собственных значений, оценки роста главного элемента в столбце значительно меньше, чем для общих матриц.

Что касается ортогональной триангуляризации, то метод Хаусхолдера лучше метода Гивенса по скорости, а также, вообще говоря, и по точности, за исключением, возможно, вычислений с плавающей запятой без накопления. Численная устойчивость обоих методов безусловно гарантируется тем, что мы имеем очень хорошие априорные оценки для эквивалентных возмущений исходной матрицы. За исключением ошибок округления, 2-норма каждого из столбцов остается постоянной, поэтому во время преобразования у нас, по-видимому, не будет какого-либо опасного роста величины элементов.

58. Если подходить к решению систем несколько осторожно и иметь быстродействующую вычислительную машину, то можно склониться к использованию метода Хаусхолдера для общей триангуляризации. На практике я предпочитал использовать метод Гаусса с выбором главного элемента по столбцу, рискуя маловероятной возможностью большого роста опорных элементов. Искушение применять его было большое, потому что вычислительные машины, которыми я пользовался, имели устройства для накопления скалярных произведений. Результаты, полученные таким путем, вероятно, в общем случае были *более* точными, чем результаты, которые можно было бы получить по методу Хаусхолдера. Для матриц Хессенберга даже самые осторожные вычислители могут, наверное, отдать предпочтение методу Гаусса с выбором главного элемента по столбцу; для трехдиагональных матриц вообще нет больше никаких причин для того, чтобы отдать предпочтение методу Хаусхолдера.

Если целью триангуляризации является вычисление n главных миноров, то методы (I) и (III) исключаются. Мы оказываемся перед выбором между методом Гивенса и использованием матриц типа M'_{ji} . Первый обладает гарантированной устойчивостью, но требует в четыре раза больше умножений, чем второй. Опять я отдавал предпочтение на практике второму, так как случаи неустойчивости в действительности довольно редки. Заметим, что в каждом из этих методов нельзя использовать преимущества накопления скалярных произведений. Более осторожные

все же могли бы предпочесть метод Гивенса, за исключением, возможно, матриц Хессенберга и трехдиагональных матриц, однако имеет смысл заметить, что метод (II) может быть выполнен с двойной точностью примерно за то же самое время, что и метод Гивенса с одинарной, и поэтому почти наверняка даст более высокую точность.

На протяжении этой книги время от времени будем обращаться к выбору преобразований, которые только что описали. Вероятно, можно сказать, что ортогональные преобразования имеют меньше преимуществ по сравнению с другими элементарными преобразованиями для данных задач, чем для тех, которые будут рассмотрены позднее.

Обратная подстановка

59. Читатель может удивиться, что мы сравниваем методы триангуляризации до рассмотрения ошибок, полученных в обратной подстановке. Мы сделали это по двум причинам. Во-первых, если мы интересуемся вычислением определителя, то обратной подстановки нет. Во-вторых, ошибки, полученные в обратной подстановке, значительно менее существенны, чем можно было бы ожидать.

Ограничимся довольно кратким анализом, так как эта задача подробно исследована Уилкинсоном (1963b, стр. 99—107). Рассмотрим решение системы уравнений

$$Lx = b \quad (59.1)$$

с нижней треугольной матрицей коэффициентов L (не обязательно с единичными диагональными элементами). Решение системы с верхней треугольной матрицей аналогично. Предположим, что $x_1, x_2, \dots, x_{r-1}$ вычислены из первых $(r-1)$ уравнений (59.1) с использованием плавающей запятой с накоплением. Тогда

$$\begin{aligned} x_r &= f l_2 [(-l_{r1}x_1 - l_{r2}x_2 - \dots - l_{r,r-1}x_{r-1} + b_r)/l_{rr}] \equiv \\ &\equiv [-l_{r1}x_1(1 + \varepsilon_1) - l_{r2}x_2(1 + \varepsilon_2) - \dots - l_{r,r-1}x_{r-1}(1 + \varepsilon_{r-1}) + \\ &\quad + b_r(1 + \varepsilon_r)](1 + \varepsilon)/l_{rr}, \end{aligned} \quad (59.2)$$

где (гл. 3, § 9)

$$|\varepsilon_i| < \frac{3}{2}(r-i+2)2^{-2t_2}, \quad |\varepsilon_r| \leq \frac{3}{2}2^{-2t}, \quad |\varepsilon| \leq 2^{-t}. \quad (59.3)$$

Если разделить числитель и знаменатель на $(1 + \varepsilon_r)$, то получим

$$\begin{aligned} x_r &= [-l_{r1}x_1(1 + \eta_1) - l_{r2}x_2(1 + \eta_2) - \dots \\ &\quad \dots - l_{r,r-1}x_{r-1}(1 + \eta_{r-1}) + b_r]/l_{rr}(1 + \eta), \end{aligned} \quad (59.4)$$

где

$$|\eta_i| \leq \frac{3}{2}(r-i+3)2^{-2t_2}, \quad |\eta| < 2^{-t}(1,00001). \quad (59.5)$$

Умножая уравнение (59.4) на $l_{rr}(1 + \eta)$ и перегруппировывая, находим

$$\dots + l_{r,r-1}(1 + \eta_{r-1})x_{r-1} + l_{rr}(1 + \eta)x_r = b_r, \quad (59.6)$$

показывая, что вычисленное x_i точно удовлетворяет уравнению с коэффициентами $l_{ri}(1 + \eta_i)$ ($i = 1, \dots, r-1$) и $l_{rr}(1 + \eta)$. Относительное возмущение во всех коэффициентах порядка 2^{-2t_2} , кроме l_{rr} , который может иметь относительное возмущение порядка 2^{-t} . Поэтому вычисленный вектор x удовлетворяет системе

$$(L + \delta L)x = b, \quad (59.7)$$

где δL есть функция от b , но в любом случае она равномерно ограничена сверху, как показано в (59.8) для $n = 5$.

$$\|\delta L\| \leq 2^{-t}(1,00001) \begin{bmatrix} |l_{11}| & & & & \\ & |l_{22}| & & & \\ & & |l_{33}| & & \\ & & & |l_{44}| & \\ & & & & |l_{55}| \end{bmatrix} +$$

$$+ \frac{3}{2} 2^{-2t_2} \begin{bmatrix} 0 & & & & \\ 4|l_{21}| & 0 & & & \\ 5|l_{31}| & 4|l_{32}| & 0 & & \\ 6|l_{41}| & 5|l_{42}| & 4|l_{43}| & 0 & \\ 7|l_{51}| & 6|l_{52}| & 5|l_{53}| & 4|l_{54}| & 0 \end{bmatrix}. \quad (59.8)$$

60. Это одна из самых хороших оценок ошибок, полученных в матричном анализе. Насколько она хороша, станет ясно, когда рассмотрим вектор-невязку. Имеем

$$b - Lx = \delta Lx, \quad (60.1)$$

откуда

$$\|b - Lx\| \leq \|\delta L\| \|x\|. \quad (60.2)$$

Если элементы L ограничены единицей, то

$$\|\delta L\| < 2^{-t}(1,00001) + \frac{3}{4} 2^{-2t_2}(n+3)^2 \quad (60.3)$$

для 1-, 2- или ∞ -нормы. Если $n^2 2^{-t} \ll 1$, то второй член незначителен. Рассмотрим теперь вектор $\bar{x}$, полученный округлением точного решения x_e до t разрядов в арифметике с плавающей запятой. Можем написать

$$\bar{x} = x_e + d. \quad (60.4)$$

Очевидно, что

$$\|d\|_{\infty} \leq 2^{-t} \|x_e\|_{\infty}, \quad (60.5)$$

и, следовательно,

$$b - L\bar{x} = b - L(x_e + d) = -Ld, \quad (60.6)$$

откуда

$$\|b - L\bar{x}\|_{\infty} = \|Ld\|_{\infty} \leq \|L\|_{\infty} \|d\|_{\infty} \leq n 2^{-t} \|x\|_{\infty}. \quad (60.7)$$

Нетрудно построить примеры, для которых эта оценка достижима. Это означает, что можно ожидать, что невязки, соответствующие вычисленному решению треугольной системы уравнений, будут меньше, чем невязки, соответствующие правильно округленному истинному решению.

Высокая точность вычисленного решения треугольной системы уравнений

61. В последнем параграфе мы дали оценку невязок вычисленных решений. Как ни замечательна эта оценка, она недооценивает точности самого решения, когда L очень плохо обусловлена. Мы показали, что вычисленное x удовлетворяет (59.7), и дали оценку (59.8) для δL . Все, что мы можем получить из нее, так это

$$x = (L + \delta L)^{-1} b = (I + L^{-1} \delta L)^{-1} L^{-1} b \quad (61.1)$$

и, следовательно,

$$\|x - L^{-1} b\| / \|L^{-1} b\| \leq \|L^{-1} \delta L\| / (1 - \|L^{-1} \delta L\|) \quad (61.2)$$

при условии $\|L^{-1} \delta L\| < 1$. До сих пор, когда мы имели оценку такого вида, мы заменяли $\|L^{-1} \delta L\|$ на $\|L^{-1}\| \|\delta L\|$, и обычно это не приводило к большим переоценкам. Однако для нашего частного случая это не так. Действительно, пренебрегая членом 2^{-2t_2} , получаем

$$\|L^{-1} \delta L\| \leq \|L^{-1}\| \|\delta L\| \leq 2^{-t} [\|L^{-1}\| \max_i |l_{ii}|]. \quad (61.3)$$

Здесь диагональные элементы L^{-1} равны $1/l_{ii}$ и, следовательно, матрица в квадратных скобках правой части (61.3) есть нижняя треугольная матрица с единичными диагональными элементами. Предположим, что элементы L^{-1} удовлетворяют соотношениям

$$|(L^{-1})_{ij}| \leq K |(L^{-1})_{jj}| \quad (i > j) \quad (61.4)$$

для некоторого K . Тогда (61.3) дает

$$\|L^{-1} \delta L\| \leq 2^{-t} K T, \quad (61.5)$$

где T есть нижняя треугольная матрица, каждый элемент которой равен единице. Следовательно,

$$\|L^{-1} \delta L\|_{\infty} \leq 2^{-t} n K. \quad (61.6)$$

Аналогично, если мы решаем систему уравнений

$$Ux = b \quad (61.7)$$

с верхней треугольной матрицей, то вычисленное решение удовлетворяет соотношению

$$(U + \delta U)x = b, \quad (61.8)$$

где, с точностью до членов порядка 2^{-2t_2} , δU есть такая диагональная матрица, что

$$|(\delta U)_{ii}| < 2^{-t} (1,00001) |u_{ii}|. \quad (61.9)$$

Снова, если

$$|(U^{-1})_{ij}| \leq K |(U^{-1})_{jj}| \quad (i < j), \quad (61.10)$$

то

$$\|U^{-1} \delta U\|_{\infty} \leq 2^{-t} n K. \quad (61.11)$$

Для многих очень плохо обусловленных треугольных матриц K порядка единицы. Например, такой матрицей является L^T из § 45. Для этой верхней треугольной матрицы мы можем взять $K = 2,5$, поэтому

можно предполагать, что вычисленная обратная к L^T матрица будет близка к точной обратной, несмотря на тот факт, что L^T очень плохо обусловлена. В действительности максимальная ошибка в любой компоненте меньше, чем одна единица последнего знака наибольшего элемента. Интересно заметить, что для соответствующей нижней треугольной матрицы L значение K существенно больше.

62. Для того чтобы подчеркнуть, что довольно неожиданная высокая точность имеет место не только у матриц со свойством (61.4), рассмотрим матрицы, имеющие почти противоположное свойство. Предположим, что L — нижняя треугольная матрица и имеет положительные диагональные элементы и отрицательные недиагональные. Покажем, что если b есть правая часть с неотрицательными элементами, то *вычисленное решение имеет малую относительную ошибку, хотя L может быть очень плохо обусловленной*.

Обозначим точное решение через y , а вычисленное через x и определим E соотношением

$$x_r = y_r (1 + E_r). \quad (62.1)$$

Покажем, что

$$(1 - 2^{-t})^r \left(1 - \frac{3}{2} 2^{-2t}\right)^{r(r+1)/2} \leq 1 + E_r \leq (1 + 2^{-t})^r \left(1 + \frac{3}{2} 2^{-2t}\right)^{r(r+1)/2}, \quad (62.2)$$

причем эту форму оценки ошибки удобно сохранить на некоторое время. Доказательство по индукции. Предположим, что оно верно до x_{r-1} . Теперь x_r дается формулой (59.2), где из гл. 3 § 8 можем заменить оценки (59.3) на более точные:

$$\left. \begin{aligned} \left(1 - \frac{3}{2} 2^{-2t}\right)^r &\leq 1 + \varepsilon_1 \leq \left(1 + \frac{3}{2} 2^{-2t}\right)^r, \\ \left(1 - \frac{3}{2} 2^{-2t}\right)^{r-i+2} &\leq 1 + \varepsilon_i \leq \left(1 + \frac{3}{2} 2^{-2t}\right)^{r-i+2}, \\ 1 - 2^{-t} &\leq 1 + \varepsilon \leq 1 + 2^{-t}. \end{aligned} \right\} \quad (62.3)$$

Вспоминая, что l_{ri} ($i < r$) отрицательные, а x_i и y_i , очевидно, неотрицательны, выражение для x_r будет взвешенным средним значением неотрицательных членов, и из нашего индуктивного предположения следует, что E_r удовлетворяет оценкам (62.2).

Легко построить матрицы этого вида, имеющие сколь угодно большие числа обусловленности. Заметим, что для того, чтобы обратить L , мы берем n правых частей от e_1 до e_n , каждая из которых состоит целиком из неотрицательных элементов. Следовательно, все элементы вычисленной обратной матрицы имеют малые относительные ошибки. Читатель может почувствовать, что аргументы, которые мы только что привели, весьма тенденциозны. Признавая, что это верно, все же имеет смысл подчеркнуть, что вычисленные решения треугольных систем, вообще говоря, намного точнее, чем это можно установить из оценки, полученной для вектора невязок, и эта высокая точность часто важна на практике.

Мы исследовали лишь случай, когда используется накопление с плавающей запятой. Для других видов вычислений полученные оценки несколько слабее, но в общих чертах сохраняют те же самые свойства. Подробный анализ был дан Уилкинсоном (1961b и 1963b, стр. 99—107).

Решение общей системы уравнений

63. Теперь вернемся к решению общей системы уравнений

$$Ax = b. \quad (63.1)$$

Предположим, что A была приведена по некоторому методу к треугольному виду, и мы имеем матрицы L и U такие, что

$$LU = A + F, \quad (63.2)$$

где оценка для F зависит от особенностей триангуляризации. Если использовался некоторый выбор главного элемента, то A означает исходную матрицу с соответственно переставленными строками. Чтобы найти решение, нужно решить две треугольные системы уравнений

$$Ly = b \quad \text{и} \quad Ux = y. \quad (63.3)$$

Если мы используем x и y для обозначения вычисленных решений, то для плавающей арифметики с накоплением

$$(L + \delta L)y = b \quad \text{и} \quad (U + \delta U)x = y \quad (63.4)$$

с оценками для δL , данными в (59.8), и соответствующими оценками для δU . Следовательно, вычисленное x удовлетворяет системе

$$(L + \delta L)(U + \delta U)x = b, \quad (63.5)$$

т. е.

$$(A + G)x = b, \quad (63.6)$$

где

$$G = F + \delta LU + L\delta U + \delta L\delta U. \quad (63.7)$$

Если мы предположим, что $|a_{ij}| \leq 1$, что использовался выбор главного элемента по столбцу и что все элементы U оставались по модулю меньше единицы, то для вычислений с плавающей запятой и накоплением мы, например, имеем

$$\|G\|_{\infty} \leq n2^{-t} + (1,00001)2^{-t}n + (1,00001)2^{-t}n + O(n^32^{-2t}). \quad (63.8)$$

Следовательно, при условии, что n^22^{-t} существенно меньше единицы, оценка $\|G\|_{\infty}$ приближенно равна $3 \cdot 2^{-t}n$. Заметим, что одна треть этой оценки получается от ошибок триангуляризации, а две другие трети от ошибок при решении треугольных систем уравнений.

Из (63.6) находим

$$r = b - Ax = Gx, \quad \|r\|_{\infty} \leq \|G\|_{\infty} \|x\|_{\infty} < (3,1)2^{-t}n \|x\|_{\infty}, \quad (63.9)$$

причем последняя оценка имеет место, когда, например, $n^22^{-t} < 0,01$. Итак, мы имеем оценку невязки, которая зависит только от *величины вычисленного решения, а не от его точности*. Если A очень плохо обусловлена, то обычно вычисленное решение будет очень неточным, но тем не менее, если $\|x\|_{\infty}$ порядка единицы, то невязка будет самое большее порядка $n2^{-t}$. Это означает, что вычисленное x будет точным решением системы $Ax = b + \delta b$ для некоторого δb , имеющего норму порядка не больше чем $n2^{-t}$. Опять, как и в § 60, поучительно сравнить эту невязку с невязкой, соответствующей правильно округленному решению. Верхняя оценка для нормы этой невязки равна $n2^{-t} \|x\|_{\infty}$, так что оценка из (63.9) соответствующая вычисленному решению, лишь в три раза больше.

Вычисление общей обратной матрицы

64. Беря b последовательно равными $e_1, e_2, \dots, e_n$, получим n столбцов матрицы, обратной к A . Для r -го столбца x_r имеем

$$(L + \delta L_r)(U + \delta U_r)x_r = e_r, \quad (64.1)$$

где мы написали δL_r и δU_r для того, чтобы подчеркнуть, что эквивалентные возмущения различны для каждой правой части, хотя они равномерно ограничены сверху оценкой (59.8). Записывая

$$AX - I = K \quad (64.2)$$

и используя равномерные оценки для δL_r и δU_r , находим, что

$$\|K\|_\infty \leq (3,1)n2^{-t}\|X\|_\infty, \quad \text{если} \quad n^2 2^{-t} < 0,01. \quad (64.3)$$

Если $\|K\|_\infty$ меньше единицы, то (64.2) означает, что ни A , ни X невырожденные. В этом случае имеем

$$X - A^{-1} = A^{-1}K, \quad (64.4)$$

откуда

$$\|X - A^{-1}\|_\infty \leq (3,1)n2^{-t}\|A^{-1}\|_\infty\|X\|_\infty \quad (64.5)$$

и

$$\|X\|_\infty \leq \|A^{-1}\|_\infty/[1 - (3,1)n2^{-t}\|A^{-1}\|_\infty]. \quad (64.6)$$

Теперь из (64.5)

$$\|X - A^{-1}\|_\infty/\|A^{-1}\|_\infty \leq (3,1)n2^{-t}\|A^{-1}\|_\infty/[1 - (3,1)n2^{-t}\|A^{-1}\|_\infty]. \quad (64.7)$$

Из этих результатов заключаем следующее. Если мы вычисляем обратную матрицу в плавающей арифметике с накоплением, используя треугольное разложение с перестановками, и если ни один из элементов U не превосходит единицы, то при условии, что вычисленная X такова, что $(3,1)n2^{-t}\|X\|_\infty$ меньше единицы, соотношения (64.6) и (64.7) выполняются. Если, в частности, эта величина мала, то (64.5) показывает, что вычисленная X имеет малую относительную ошибку. И наоборот, любая нормированная матрица, для которой $(3,1)n2^{-t}\|A^{-1}\|_\infty < 1$, должна дать вычисленную X , удовлетворяющую (64.7) при условии, что не было роста ведущих элементов.

К сожалению, должно быть сделано замечание относительно роста ведущих элементов, но так как он будет редко, этот факт на практике не очень существен.

Для других, менее точных способов вычислений мы имеем соответствующие оценки, которые, конечно, несколько слабее, но вследствие статистического распределения ошибок, оценки, которые мы только что дали, на практике не превышаются.

Точность вычисленных решений

65. Заметим, что для обращения матрицы мы получили много лучшие результаты, чем результаты, которые были получены для решения систем линейных уравнений. Это потому, что малая матрица-невязка $(AX - I)$ обязательно означает, что X есть достаточно хорошая обратная матрица, тогда как малый вектор-невязка $(Ax - b)$ не обязательно означает, что

x есть хорошее решение. На практике, когда мы вычислили обратную матрицу по любому методу, мы можем вычислить $(AX - I)$ и поэтому получить надежную оценку ошибки в X .

Естественно спросить, нет ли какого-нибудь аналогичного способа оценки точности вычисленного решения системы $Ax = b$. Если мы напишем

$$b - Ax = r, \quad (65.1)$$

то ошибка в x равна $A^{-1}r$ и, следовательно, если мы имеем оценку для $\|A^{-1}\|$, то можем получить оценку и для ошибки. Однако эта оценка может быть очень неутешительной даже тогда, когда наша оценка для $\|A^{-1}\|$ очень точная. Предположим, например, что A есть такая нормированная матрица, что $\|A^{-1}\|_{\infty} = 10^8$, а b такая нормированная правая часть, что $\|A^{-1}b\|_{\infty} = 1$. Если x есть вектор, полученный округлением $A^{-1}b$ до десяти десятичных знаков, то

$$\|r\|_{\infty} = \|b - Ax\|_{\infty} \leq \frac{1}{2} n 10^{-10} \quad (65.2)$$

и эта оценка вполне реалистична. Следовательно, если нам дан такой x и точная оценка 10^8 для $\|A^{-1}\|_{\infty}$, то мы могли бы лишь показать, что норма ошибки в x ограничена величиной $\frac{1}{2} n 10^{-2}$, а это очень неутешительный результат. Для того чтобы суметь определить точность такого решения, нужна достаточно точная приближенная обратная матрица X и достаточно точная оценка для $\|X - A^{-1}\|$.

Плохо обусловленные матрицы, которые дают немалые ведущие элементы

66. Можно было бы думать, что если нормированная матрица A очень плохо обусловлена, то при треугольном разложении с выбором главного элемента это должно обнаружиться появлением очень малого ведущего элемента. Можно рассуждать, что если A — вырожденная и если разложение было выполнено точно, то один из ведущих элементов обязательно должен быть нулевым, и следовательно, если A плохо обусловлена, то один из ведущих элементов должен быть мал.

К сожалению, это совсем неверно даже тогда, когда A — симметричная (это был бы благоприятный случай, так как все s_i равны единице). Рассмотрим матрицу, имеющую собственные значения

$$1,00, 0,95, \dots, 0,05 \text{ и } 10^{-10}.$$

Такая матрица почти вырождена в том смысле, что $(A - 10^{-10}I)$ — точно вырожденная. Но произведение p из ведущих элементов равно произведению собственных значений, и поэтому

$$p = 10^{-10} (0,05)^{20} 20!$$

Если все ведущие элементы были равны, то каждый из них приблизительно равен 0,1, и мы, конечно, не можем рассматривать их как патологически малые. Может показаться, что это доказательство искусственно, но в гл. 5, § 56 мы увидим, что такие примеры могут действительно появиться. Это явление может возникнуть не только в методе Гаусса с выбором главного элемента по столбцу, но и в методах Гивенса и Хаусхолдера.

Есть ряд признаков, которые могут появиться во время решения методом Гаусса с выбором главного элемента по столбцу системы уравнений $Ax = b$ (с нормированными A и b) и которые сразу укажут на ее плохую обусловленность. Это появление

- (i) «малого» ведущего элемента,
- (ii) «большого» численного решения,
- (iii) «большого» вектора-невязки.

К сожалению, можно построить патологически плохо обусловленные системы уравнений, у которых нет ни одного из этих признаков. Работая с t -разрядной арифметикой, мы могли бы вполне получить решение, имеющее элементы порядка единицы и поэтому обязательно имеющее невязку порядка 2^{-t} , но которое тем не менее было бы неверно даже в своих старших разрядах.

Итерационное улучшение приближенного решения

67. Рассуждения предыдущих параграфов показывают, что получить надежную оценку ошибки вычисленного решения, которая к тому же была бы удовлетворительна для достаточно точных решений плохо обусловленных уравнений, совсем не просто. Теперь покажем, как улучшить приближенное решение и в то же время получить надежную информацию о точности исходного и улучшенного решения.

Предположим, что мы получили приближенные треугольные матрицы L и U такие, что

$$LU = A + F. \quad (67.1)$$

Мы можем использовать их для получения последовательности приближенных решений $x^{(n)}$, сходящихся к точному решению x системы $Ax = b$, с помощью такого итерационного процесса:

$$\left. \begin{aligned} x^{(0)} &= 0, \quad r^{(0)} = b, \\ LUd^{(k-1)} &= r^{(k-1)}, \quad x^{(k)} = x^{(k-1)} + d^{(k-1)}, \quad r^{(k)} = b - Ax^{(k)}. \end{aligned} \right\} \quad (67.2)$$

Если бы этот процесс мог быть выполнен без ошибок округления, то мы имели бы

$$x^{(k)} = x^{(k-1)} + (A + F)^{-1} (b - Ax^{(k-1)}) = x^{(k-1)} + (A + F)^{-1} A (x - x^{(k-1)}), \quad (67.3)$$

откуда

$$x - x^{(k)} = [I - (A + F)^{-1} A] (x - x^{(k-1)}) = [I - (A + F)^{-1} A]^k x. \quad (67.4)$$

Поэтому достаточное условие для сходимости точного итерационного процесса таково:

$$\|I - (A + F)^{-1} A\| < 1, \quad (67.5)$$

и оно заведомо выполнено, если

$$\|A^{-1}\| \|F\| < \frac{1}{2}. \quad (67.6)$$

Аналогично

$$r^{(k)} = A [I - (A + F)^{-1} A]^k x, \quad (67.7)$$

и невязки также сходятся к нулю, если (67.6) выполнено.

Если A разложена на треугольные множители по методу Гаусса с выбором главного элемента по столбцу с использованием плавающей арифметики и накопления, то, вообще говоря,

$$\|F\|_{\infty} \leq n2^{-t} \quad (67.8)$$

и, следовательно, точный итерационный процесс будет сходиться, если

$$\|A^{-1}\|_{\infty} < 2^{t-1}/n. \quad (67.9)$$

Если $n2^{-t} \|A^{-1}\|_{\infty} < 2^{-p}$, то $x^{(k)}$ будет получать по крайней мере по p двоичных верных знаков за итерацию, а $r^{(k)}$ будет уменьшаться по крайней мере в 2^{-p} раза. Если $n2^{-t} \|A^{-1}\|_{\infty}$ значительно больше, чем $1/2$, то нельзя, вообще говоря, ожидать сходимости процесса.

Влияние ошибок округления на итерационный процесс

68. В действительности мы не можем точно осуществить итерационный процесс, и на первый взгляд кажется, что влияние ошибок округления может быть губительным. Если итерация выполняется, используя t -разрядную арифметику любого вида, то точность $x^{(k)}$ не может неограниченно увеличиваться с ростом k , так как мы используем лишь t разрядов для представления компонент.

Это, конечно, верно, но когда мы обращаемся к невязкам, то находим довольно странную ситуацию. Мы показали в § 63, что если, например, использовать плавающую арифметику с накоплением, то невязка, соответствующая первому приближению, почти наверняка должна быть того же порядка, что и невязка, соответствующая правильно округленному решению. Поэтому мы едва ли сможем ожидать, что невязка будет уменьшаться в 2^{-p} раз с каждой итерацией, и априори кажется невероятно, что точность последовательных приближений будет устойчиво улучшаться, если невязки остаются приблизительно постоянными. Тем не менее это именно так, и чтобы практический процесс был близок к точному процессу, необходимо при вычислении невязок накапливать скалярные произведения.

Уилкинсон (1963b, стр. 121—126) дал подробный анализ ошибок итерационного процесса в арифметике с фиксированной запятой, используя накопления на всех соответствующих этапах и вычисляя каждый $x^{(k)}$ и $d^{(k)}$ как блочно-плавающие векторы. Не будем повторять здесь анализа, но так как сам метод важен для последующих глав, опишем его подробно. Соответствующие изменения при использовании плавающего накопления очевидны, но этот процесс в действительности несколько проще.

Итерационный процесс в вычислении с фиксированной запятой

69. Предположим, что элементы A заданы с одинарной точностью в арифметике с фиксированной запятой, а L и U вычислены при треугольном разложении с перестановками с использованием накопления скалярных произведений. Предположим далее, что ни один из элементов U не превосходит по модулю единицу. Элементы правой части b также являются числами с фиксированной запятой, но могут быть заданы как с одинарной, так и с двойной точностью.

Мы представим k -е вычисленное приближение $x^{(k)}$ в виде $2^p k y^{(k)}$, где $y^{(k)}$ — нормированный блочно-плавающий вектор. В общем случае p_k

будет постоянным для всех итераций. (Это верно, по крайней мере, если A не очень плохо обусловлена; однако если есть очень близкие округления в наибольшей компоненте, то p_k может изменяться на ± 1 от итерации к итерации.) Теперь общий шаг определяется следующим образом:

(i) Вычисляем точную невязку $r^{(k)}$,

$$r^{(k)} = b - 2^{p_k} A y^{(k)},$$

получая сначала этот вектор как ненормализованный блочно-плавающий вектор с двойной точностью. Ясно, что на этом этапе мы можем использовать вектор b с двойной точностью, не требуя каких-либо умножений с двойной точностью.

(ii) Нормализуем невязку $r^{(k)}$ и затем округляем ее для получения блочно-плавающего вектора с одинарной точностью, который обозначим через $2^{q_k} s^{(k)}$.

(iii) Решаем уравнение $LUx = s^{(k)}$, получая решение в виде блочно-плавающего вектора с одинарной точностью. Подбираем нормализующий множитель, чтобы получить вычисленное решение $d^{(k)}$ системы $LUx = 2^{q_k} s^{(k)}$.

(iv) Вычисляем следующее приближение:

$$x^{(k+1)} = x^{(k)} + d^{(k)},$$

получая $x^{(k+1)}$ как блочно-плавающий вектор с одинарной точностью.

Простой пример итерационного процесса

70. В табл. 6 мы даем пример применения этого метода к матрице третьего порядка, где вычисления выполнялись с использованием шестизначной арифметики с фиксированной запятой и накоплением скалярных произведений. Матрица имеет число обусловленности между 10^4 и 10^5 .

Таблица 6

*Итерационное решение системы уравнений
с тремя различными правыми частями*

A			b_1	b_2	b_3
0,876543	0,617341	0,589973	0,863257	0,8632572136	0,4135762133
0,612314	0,784461	0,827742	0,820647	0,8206474986	0,3712631241
0,317321	0,446779	0,476349	0,450098	0,4500984014	0,5163213142
L			U		
1,000000			0,876543	0,617341	0,589973
0,698556	1,000000			0,353214	0,415613
0,362014	0,632175	1,000000			0,000030

Итерации, соответствующие первой правой части

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
0,639129	0,636280	0,636330	0,636329	0,636329
-0,050678	-0,029140	-0,029513	-0,029507	-0,029507
0,566667	0,548363	0,548680	0,548674	0,548674
$10^6 r^{(1)}$	$10^6 r^{(2)}$	$10^6 r^{(3)}$	$10^6 r^{(4)}$	
0,326160	0,172501	-0,407897	0,304438	
0,204138	-0,044726	-0,450687	0,421313	
-0,446030	-0,032507	-0,252623	0,242118	

Если A разложена на треугольные множители по методу Гаусса с выбором главного элемента по столбцу с использованием плавающей арифметики и накопления, то, вообще говоря,

$$\|F\|_{\infty} \leq n2^{-t} \quad (67.8)$$

и, следовательно, точный итерационный процесс будет сходиться, если

$$\|A^{-1}\|_{\infty} < 2^{t-1}/n. \quad (67.9)$$

Если $n2^{-t} \|A^{-1}\|_{\infty} < 2^{-p}$, то $x^{(k)}$ будет получать по крайней мере по p двоичных верных знаков за итерацию, а $r^{(k)}$ будет уменьшаться по крайней мере в 2^{-p} раза. Если $n2^{-t} \|A^{-1}\|_{\infty}$ значительно больше, чем $1/2$, то нельзя, вообще говоря, ожидать сходимости процесса.

Влияние ошибок округления на итерационный процесс

68. В действительности мы не можем точно осуществить итерационный процесс, и на первый взгляд кажется, что влияние ошибок округления может быть губительным. Если итерация выполняется, используя t -разрядную арифметику любого вида, то точность $x^{(k)}$ не может неограниченно увеличиваться с ростом k , так как мы используем лишь t разрядов для представления компонент.

Это, конечно, верно, но когда мы обращаемся к невязкам, то находим довольно странную ситуацию. Мы показали в § 63, что если, например, использовать плавающую арифметику с накоплением, то невязка, соответствующая первому приближению, *почти наверняка должна быть того же порядка, что и невязка, соответствующая правильно округленному решению*. Поэтому мы едва ли сможем ожидать, что невязка будет уменьшаться в 2^{-p} раз с каждой итерацией, и априори кажется невероятно, что точность последовательных приближений будет устойчиво улучшаться, если невязки остаются приблизительно постоянными. Тем не менее это именно так, и чтобы практический процесс был близок к точному процессу, необходимо при вычислении невязок накапливать скалярные произведения.

Уилкинсон (1963b, стр. 121—126) дал подробный анализ ошибок итерационного процесса в арифметике с фиксированной запятой, используя накопления на всех соответствующих этапах и вычисляя каждый $x^{(k)}$ и $d^{(k)}$ как блочно-плавающие векторы. Не будем повторять здесь анализа, но так как сам метод важен для последующих глав, опишем его подробно. Соответствующие изменения при использовании плавающего накопления очевидны, но этот процесс в действительности несколько проще.

Итерационный процесс в вычислении с фиксированной запятой

69. Предположим, что элементы A заданы с одинарной точностью в арифметике с фиксированной запятой, а L и U вычислены при треугольном разложении с перестановками с использованием накопления скалярных произведений. Предположим далее, что ни один из элементов U не превосходит по модулю единицу. Элементы правой части b также являются числами с фиксированной запятой, но могут быть заданы как с одинарной, так и с двойной точностью.

Мы представим k -е вычисленное приближение $x^{(k)}$ в виде $2^p y^{(k)}$, где $y^{(k)}$ — нормированный блочно-плавающий вектор. В общем случае p_k

будет постоянным для всех итераций. (Это верно, по крайней мере, если A не очень плохо обусловлена; однако если есть очень близкие округления в наибольшей компоненте, то p_k может изменяться на ± 1 от итерации к итерации.) Теперь общий шаг определяется следующим образом:

(i) Вычисляем точную невязку $r^{(k)}$,

$$r^{(k)} = b - 2^{p_k} A y^{(k)},$$

получая сначала этот вектор как ненормализованный блочно-плавающий вектор с двойной точностью. Ясно, что на этом этапе мы можем использовать вектор b с двойной точностью, не требуя каких-либо умножений с двойной точностью.

(ii) Нормализуем невязку $r^{(k)}$ и затем округляем ее для получения блочно-плавающего вектора с одинарной точностью, который обозначим через $2^{q_k} s^{(k)}$.

(iii) Решаем уравнение $LUx = s^{(k)}$, получая решение в виде блочно-плавающего вектора с одинарной точностью. Подбираем нормализующий множитель, чтобы получить вычисленное решение $d^{(k)}$ системы $LUx = 2^{q_k} s^{(k)}$.

(iv) Вычисляем следующее приближение:

$$x^{(k+1)} = x^{(k)} + d^{(k)},$$

получая $x^{(k+1)}$ как блочно-плавающий вектор с одинарной точностью.

Простой пример итерационного процесса

70. В табл. 6 мы даем пример применения этого метода к матрице третьего порядка, где вычисления выполнялись с использованием шестизначной арифметики с фиксированной запятой и накоплением скалярных произведений. Матрица имеет число обусловленности между 10^4 и 10^5 .

Таблица 6

*Итерационное решение системы уравнений
с тремя различными правыми частями*

A			b_1	b_2	b_3
[0,876543	0,617341	0,589973]	[0,863257]	[0,8632572136]	[0,4135762133]
[0,612314	0,784461	0,827742]	[0,820647]	[0,8206474986]	[0,3712631241]
[0,317321	0,446779	0,476349]	[0,450098]	[0,4500984014]	[0,5163213142]
L			U		
[1,000000			[0,876543	0,617341	0,589973]
[0,698556	1,000000			0,353214	0,415613]
[0,362014	0,632175	1,000000]			0,000030]

Итерации, соответствующие первой правой части

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
0,639129	0,636280	0,636330	0,636329	0,636329
-0,050678	-0,029140	-0,029513	-0,029507	-0,029507
0,566667	0,548363	0,548680	0,548674	0,548674
$10^6 r^{(1)}$	$10^6 r^{(2)}$	$10^6 r^{(3)}$	$10^6 r^{(4)}$	
0,326160	0,172501	-0,407897	0,304438	
0,204138	-0,044726	-0,450687	0,421343	
-0,446030	-0,032507	-0,252623	0,242118	

Продолжение табл. 6

Итерации, соответствующие второй правой части

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
0,639129	0,636815	0,636855	0,636855	0,636855
-0,050678	-0,033187	-0,033490	-0,033484	-0,033485
0,566667	0,551803	0,552060	0,552056	0,552056
$10^6 r^{(1)}$	$10^6 r^{(2)}$	$10^6 r^{(3)}$	$10^6 r^{(4)}$	
0,539760	0,307503	0,677045	-0,667109	
0,702738	0,147071	0,616500	-0,779298	
-0,044630	0,076211	0,335715	-0,439563	

Итерации, соответствующие третьей правой части

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
1632,2	1603,9	1604,4	1604,4	1604,4
-12336,6	-12123,1	-12126,8	-12126,8	-12126,8
10484,6	10303,2	10306,4	10306,3	10306,3
$10^4 r^{(1)}$	$10^4 r^{(2)}$	$10^4 r^{(3)}$	$10^4 r^{(4)}$	
-0,218436	0,031220	-0,389014	0,200959	
-0,098483	-0,113868	-0,638125	0,189617	
-0,099289	-0,073525	-0,372475	0,103874	

Уравнения были решены с тремя различными правыми частями, которые иллюстрируют ряд особенно интересных свойств.

Первая правая часть была выбрана так, чтобы, несмотря на плохую обусловленность A , точное решение было порядка единицы. Так как матрица не очень уж плохо обусловлена, первое приближение также порядка единицы. Соответственно первая невязка имеет компоненты порядка 10^{-6} , несмотря на то, что приближенное решение имеет ошибки во втором десятичном знаке. Неточность первого приближения сразу же обнаруживается, когда вычисляется второе. Изменения в первой компоненте решения при последовательных итерациях равны $-0,002849$, $0,000050$, $-0,000001$, $0,000000$, другие компоненты ведут себя аналогично. Большое относительное изменение, сделанное при первом исправлении, убедительно показывает, что матрица A плохо обусловлена и что число обусловленности, вероятно, порядка 10^4 .

В этом примере это очевидно, потому что последний элемент U равен $0,000030$, и это показывает, что элемент $(3, 3)$ в обратной матрице будет порядка $30\,000$. Невозможно построить матрицу третьего порядка, которая бы не обнаруживала свою плохую обусловленность таким образом, но для матриц более высокого порядка плохая обусловленность совсем неочевидна. Однако поведение итераций и в этих случаях дает надежные признаки обусловленности. Заметим, что все невязки имеют почти одну и ту же величину и в действительности последняя невязка, соответствующая правильно округленному решению, является одной из самых больших.

Вторая правая часть отличается от первой лишь в последних знаках, начиная с седьмого, и поэтому первое приближение совпадает с первым приближением, соответствующим предыдущей правой части. Однако мы можем использовать все знаки b_2 при вычислении невязки и, следовательно, после первой итерации ожидать появления дополнительных знаков. Хотя разность между двумя правыми частями порядка 10^{-6} , два решения могут отличаться на вектор порядка 10^{-2} , так как $\|A^{-1}\|$ порядка 10^4 . Это действительно оказывается так, и поэтому использование полной правой части имеет большое значение.

Третья правая часть дает решение, имеющее компоненты порядка 10^4 . Появление такого решения сразу же говорит о плохой обусловленности матрицы. Так как приближения получаются как блочно-плавающие векторы, мы имеем лишь один знак после запятой, и даже правильно округленные последние приближения могут содержать ошибки до $\pm 0,05$. Теперь на всех этапах невязки порядка 0,01, несмотря на то, что компоненты первого приближения имеют ошибки больше 100. Случайные ошибки такой величины могли бы дать невязки порядка 100.

Общие замечания по итерационному процессу

71. Картина, описанная в нашем примере, типична для матриц с числом обусловленности порядка 10^4 , и читателю рекомендуется тщательно ее изучить. Для матриц третьего порядка едва ли стоит говорить о «совокупном влиянии» ошибок округления и *важно понять, что в общем случае это накопление играет сравнительно малую роль*. Мы проиллюстрируем это показом грубых оценок различных ошибок, возникающих при решении этим методом нормированной системы уравнений с нормированной правой частью. В этих оценках учтен статистический эффект. Предполагаем, что число обусловленности порядка 2^a , максимальная компонента точного решения порядка 2^b и $n2^{a-t}$ существенно меньше $1/2$.

(а) Максимальная компонента первого и всех последующих приближений порядка 2^b .

(б) Общий уровень компонент невязки, соответствующей первому приближению, порядка $(2n)^{1/2}2^{b-t}$.

(в) Общий уровень ошибок в компонентах первого приближения порядка $n^{1/2}2^{b-t}2^a$.

(г) Общий уровень компонент невязки, соответствующей второму приближению (и всем следующим приближениям) порядка $n^{1/2}2^{b-t}$.

(д) Общий уровень ошибок во втором приближении порядка $n2^{b-2t}2^{2a}$.

Заметим, что оценки невязок содержат множитель 2^b и, следовательно, на всех этапах пропорциональны величине вычисленного решения и не зависят от числа обусловленности 2^a . Однако мы знаем, что 2^b не может быть больше 2^a , так что обусловленность матрицы определяет максимальную величину, достигаемую приближениями и, следовательно, невязками.

Хотя мы говорили о процессе как итерационном, мы ожидаем, что в общем случае должно быть очень мало итераций. Предположим, например, что мы имеем матрицу, для которой $n = 2^{12}$ и $a = 2^0$. Это соответствует очень плохой обусловленности. Тем не менее на АСЕ мы могли бы

ожидать получения на каждом шаге $48 - \left(\frac{1}{2} \cdot 12\right) = 20$ двоичных знаков, так что после второго приближения ответ должен быть верным с рабочей точностью. Если число итераций невелико, то итерационный процесс более экономичен, чем вычисление решения целиком в арифметике с двойной точностью; кроме того, он дает действительную гарантию того, что окончательный вектор совпадает с правильно округленным решением. Однако, если $n2^{-t} \|A^{-1}\| > 1/2$, то итерационный процесс, вообще говоря, не будет сходиться и ни одно из приближенных решений не будет иметь верных знаков. При работе с двойной точностью мы будем иметь несколько верных знаков при условии, что $n2^{-2t} \|A^{-1}\| < 1$.

Родственные итерационные процессы

72. Процесс, который мы только что описали, будет работать и с другими, менее точными видами арифметики, если при вычислении невязки накапливаются скалярные произведения. Если это невыгодно, то невязки должны вычисляться в арифметике с двойной точностью. Накопление скалярных произведений несущественно на любом другом этапе процесса. В общем случае, если норма эквивалентного возмущения в A , соответствующего ошибкам, сделанным при решении, ограничена величиной $n^k 2^{-l}$, то процесс будет работать при условии, что $n^k 2^{-l} \|A^{-1}\|$ меньше, чем $1/2$. На практике из-за статистического распределения ошибок процесс обычно будет работать при условии, что $n^{k/2} 2^{-l} \|A^{-1}\|$ меньше единицы. (Предполагаем здесь, что рост ведущих элементов незначителен.)

Аналогично, если мы используем триангуляризацию Гивенса или Хаусхолдера и сохраняем ортогональные преобразования, то можем выполнить такой же итерационный процесс. Опять единственным этапом, на котором важно накапливать скалярные произведения, является вычисление невязок. Для ортогональных методов триангуляризации мы можем опустить условие относительно роста ведущих элементов, так как теперь это исключается. Однако я не думаю, что это оправдывает использование ортогональных преобразований.

Если при вычислении невязок не накапливаются скалярные произведения, то, вообще говоря, итерационный процесс не будет сходиться. Действительно, тогда ошибки, сделанные при вычислении невязок, будут того же порядка, что и сами невязки (за исключением, возможно, первого этапа, когда точные невязки могут быть несколько больше для матриц большого порядка из-за множителя n^k). В общем случае последовательные итерации не будут иметь тенденции к сходимости; вся последовательность векторов будет получаться с ошибками, сравнимыми с ошибками первого приближения.

Ограниченность итерационного процесса

73. Анализ, подобный анализу, данному Уилкинсоном (1963b, стр. 121—126), показывает, что при отсутствии роста ведущих элементов любой из итерационных процессов должен сходиться для всех матриц, удовлетворяющих некоторому критерию вида

$$2^{-l} f(n) \|A^{-1}\| < 1, \quad (73.1)$$

где $f(n)$ зависит от метода триангуляризации и типа использованной арифметики. Если процесс не сходится и при триангуляризации нет роста ведущих элементов, то мы можем быть абсолютно уверенными, что A не удовлетворяет (73.1). Мы можем сказать, что матрица «слишком плохо обусловлена, чтобы систему можно было решить по рассматриваемому методу» без использования более точной арифметики.

Однако остается еще возможность, что несмотря на то, что (73.1) не выполнено, процесс явно сходится и все же дает неверный ответ. Попытки построить такие примеры сразу же показывают, насколько это маловероятно даже для одной правой части; если же нас интересуют несколько правых частей, то кажется почти непостижимо, чтобы мы могли заблуждаться на этот счет. Даже в случае плохо обусловленных матриц, о которых упоминалось в § 66 и у которых не мал ни один из ведущих элементов, решение порядка единицы, а первая невязка порядка 2^{-l} , эта плохая

обусловленность сразу же полностью обнаруживается, когда вычисляется первая поправка.

В любой автоматической программе, основанной на итерационном процессе, легко можно включить дополнительные проверки: ведущие элементы порядка $n2^{-l}$ или решения порядка $2^l/n$ являются верным признаком того, что обусловленность матрицы такова, что сходимость итераций маловероятна, а если уж она имеет место, то не гарантируется правильность решения. Даже неубывание двух последовательных поправок является безошибочным доказательством того, что (73.1) не выполняется.

Строгое обоснование итерационного метода

74. Мы всегда считали, что при включении дополнительных проверок, о которых упоминали (вероятно, даже и без них), сходимость итераций гарантирует их сходимость к точному решению и требование дальнейшего доказательства является излишним педантизмом.

Однако, если мы не хотим этого делать, мы должны получить верхнюю оценку для $\|A^{-1}\|$. Иногда такая оценка может быть получена априори из теоретических соображений. В противном случае можем использовать треугольные матрицы для того, чтобы получить вычисленную обратную матрицу X и затем получить оценку для $\|A^{-1}\|$ так же, как в § 64 с помощью вычисления $(AX - I)$. Если оценка для $\|A^{-1}\|$ обеспечивает выполнение (73.1), то мы знаем, что итерационный процесс должен сходиться к правильному решению. Иначе мы заключаем, что A слишком плохо обусловлена, для того чтобы проводить вычисления с данной точностью.

Заметим, что, имея вычисленную X , мы можем использовать другой итерационный процесс, в котором вычисляем поправку $d^{(h)}$ как произведение $Xr^{(h)}$, а не из решения $LUx = r^{(h)}$. Я анализировал этот процесс (Уилкинсон, 1963b, стр. 128—131) и показал, что он менее удовлетворителен, чем процесс из § 67, а оба процесса требуют приблизительно одного и того же числа умножений. Мы видим, что для того, чтобы получить только верхнюю оценку для $\|A^{-1}\|$, необходимо выполнить очень большую работу.

75. Существует сравнительно экономичный путь вычисления верхней оценки $\|A^{-1}\|$, хотя он может быть очень пессимистичным. Положим

$$L = D_1 + L_s, \quad U = D_2 + U_s, \quad (75.1)$$

где D_1 и D_2 — диагональные, а L_s и U_s — соответственно *строго* нижняя и верхняя треугольные матрицы. Тогда, если x есть решение уравнения

$$(|D_1| - |L_s|)(|D_2| - |U_s|)x = e, \quad (75.2)$$

где e — вектор с единичными компонентами, то

$$\|x\|_\infty \geq \|A^{-1}\|_\infty. \quad (75.3)$$

Из § 62 знаем, что вычисленное решение (75.2) имеет компоненты с малыми относительными ошибками, и, следовательно, оно дает нам надежную оценку для $\|A^{-1}\|_\infty$.

Дополнительные замечания

В этой главе мы рассмотрели лишь те аспекты линейных уравнений, с которыми непосредственно будем иметь дело в данной книге, так что даже не упоминали о решении систем итерационными методами. Для более общего ознакомления читатель

отсылается к книгам Фокса (1964) и Хаусхолдера (1964), где описаны как прямые, так и итерационные методы, и к книгам Форсайта и Вазова (1960) и Варга (1962), где описаны итерационные методы.

Задача уравнивания исследована в статье Форсайта и Страуса (1955) и позднее в статье Бауэра (1963). Эти статьи позволяют существенно понять рассматриваемую задачу, но выполнение уравнивания на практике до вычисления обратной матрицы остается трудной задачей. Уравнивание часто обсуждается в связи со стратегией выбора ведущих элементов в методе Гаусса, и это может создать впечатление, что оно не имеет отношения к тому случаю, когда используется ортогональная триангуляризация. Что это не так, становится ясно, если мы рассмотрим матрицу с элементами

$$\begin{bmatrix} 1 & 10^{10} & 10^{10} \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Первую строку следует нормировать до ортогональной триангуляризации, или же информация в других строках будет потеряна.

Первый опубликованный анализ ошибок в методах, основанных на исключении Гаусса, был анализ Неймана и Голдстейна (1947) и Тюринга (1948). Первый из них, можно сказать, стал основой современного анализа ошибок в его строгой трактовке для обращения положительно определенных матриц. Анализ, данный в этой главе, основан на работе, проделанной автором между 1957 и 1962 г., и построен в плане обратного анализа ошибок. До некоторой степени подобный анализ нескольких процессов дан Гастинелем (1960) в его диссертации.

ЭРМИТОВЫ МАТРИЦЫ

Введение

1. В этой главе опишем методы решения проблемы собственных значений для эрмитовых матриц. Мы не будем предполагать, что матрица имеет много нулевых элементов или специальна в каком-либо другом отношении. Методы, предназначенные для таких матриц, будут обсуждаться в последующих главах.

В главах 1, 2 и 3 мы видели, что проблема собственных значений для эрмитовых матриц проще, чем для матриц общего вида. Эрмитова матрица A всегда имеет линейные элементарные делители, и всегда существует унитарная матрица R такая, что

$$R^H A R = \text{diag}(\lambda_i). \quad (1.1)$$

Далее, собственные значения вещественные, и, как мы показали в гл. 2, § 31, малые возмущения в элементах A обязательно приводят к малым возмущениям в собственных значениях. Если матрица возмущений эрмитова, то собственные значения остаются вещественными (гл. 2, § 38). Хотя собственные векторы не обязательно хорошо определены, их плохое определение должно быть связано с близкими собственными значениями, так как $s_i = 1$ в (10.2) из главы 2.

2. В большей части этой главы мы ограничимся вещественными симметричными матрицами. Для таких матриц матрица R из равенства (1.1) может быть взята вещественной и поэтому ортогональной. Следует помнить, что комплексные симметричные матрицы не обладают ни одним из важнейших свойств вещественных симметричных матриц (гл. 1, § 24).

Большинство из методов решения проблемы собственных значений для матрицы A общего вида основано на выполнении ряда подобных преобразований, которые переводят A в матрицу специального вида. Для эрмитовых матриц особенно выгодно использование унитарных преобразований, так как при этом сохраняется эрмитовость со всеми присущими ей достоинствами. В методах, описанных в этой главе, мы используем элементарные унитарные преобразования, введенные в гл. 1, §§ 43–46.

Классический метод Якоби для вещественных симметричных матриц

3. В методе Якоби (1846) исходная матрица преобразуется в диагональную с помощью последовательности плоских вращений. Строго говоря, число плоских вращений, необходимых для приведения к диагональному виду, бесконечно. Это следует ожидать, так как в общем случае мы не можем найти корни полинома за конечное число шагов. На практике процесс заканчивается тогда, когда внедиагональные элементы становятся

незначительными по сравнению с рабочей точностью. Для вещественных симметричных матриц мы применяем вещественные плоские вращения.

Если мы обозначим исходную матрицу через A_0 , то процесс Якоби можно описать следующим образом. Строится последовательность матриц A_k ,

$$A_k = R_k A_{k-1} R_k^T, \quad (3.1)$$

где матрица R_k определяется по таким правилам. (Заметим, что мы использовали запись $R_k A_{k-1} R_k^T$, а не $R_k^T A_{k-1} R_k$ для того, чтобы согласовать ее с нашим общим анализом из главы 3.)

Предположим, что максимальный по модулю внедиагональный элемент A_{k-1} находится в позиции (p, q) , тогда R_k соответствует вращению в плоскости (p, q) и угол вращения θ выбирается таким образом, чтобы сделать элемент (p, q) матрицы A_k нулем.

Мы имеем

$$\left. \begin{aligned} R_{pp} &= R_{qq} = \cos \theta, & R_{pq} &= -R_{qp} = \sin \theta, \\ R_{ii} &= 1 \quad (i \neq p, q), & R_{ij} &= 0 \text{ для остальных } i, j, \end{aligned} \right\} \quad (3.2)$$

где предполагается, что p меньше q . Матрица A_k отличается от A_{k-1} лишь строками и столбцами с номерами p и q и все A_k — симметричные. Измененные элементы определяются по формулам

$$\left. \begin{aligned} a_{ip}^{(k)} &= a_{ip}^{(k-1)} \cos \theta + a_{iq}^{(k-1)} \sin \theta = a_{pi}^{(k)}, \\ a_{iq}^{(k)} &= -a_{ip}^{(k-1)} \sin \theta + a_{iq}^{(k-1)} \cos \theta = a_{qi}^{(k)}, \end{aligned} \right\} \quad (i \neq p, q) \quad (3.3)$$

$$\left. \begin{aligned} a_{pp}^{(k)} &= a_{pp}^{(k-1)} \cos^2 \theta + 2a_{pq}^{(k-1)} \cos \theta \sin \theta + a_{qq}^{(k-1)} \sin^2 \theta, \\ a_{qq}^{(k)} &= a_{pp}^{(k-1)} \sin^2 \theta - 2a_{pq}^{(k-1)} \cos \theta \sin \theta + a_{qq}^{(k-1)} \cos^2 \theta, \\ a_{pq}^{(k)} &= (a_{qq}^{(k-1)} - a_{pp}^{(k-1)}) \cos \theta \sin \theta + a_{pq}^{(k-1)} (\cos^2 \theta - \sin^2 \theta) = a_{qp}^{(k)}. \end{aligned} \right\} \quad (3.4)$$

Так как $a_{pq}^{(k)}$ должен быть нулем, то

$$\operatorname{tg} 2\theta = 2a_{pq}^{(k-1)} / (a_{pp}^{(k-1)} - a_{qq}^{(k-1)}). \quad (3.5)$$

Мы всегда будем брать θ в пределах

$$|\theta| \leq \frac{1}{4} \pi, \quad (3.6)$$

и если $a_{pp}^{(k-1)} = a_{qq}^{(k-1)}$, то берем θ равным $\pm \frac{1}{4} \pi$ в зависимости от знака $a_{pq}^{(k-1)}$.

Предполагаем, что $a_{pq}^{(k-1)} \neq 0$, так как в противном случае диагонализация уже закончена. По существу этот процесс итерационный, так как элемент, который сделан нулем при одном вращении, становится, вообще говоря, ненулевым при последующих вращениях.

Быстрота сходимости

4. Теперь докажем, что если каждый угол θ берется в пределах (3.6), то

$$A_k \rightarrow \operatorname{diag}(\lambda_i), \quad \text{когда } k \rightarrow \infty, \quad (4.1)$$

где λ_i суть собственные значения матрицы A_0 , а, следовательно, и всех матриц A_k , взятые в некотором порядке.

Обозначим

$$A_k = \operatorname{diag}(a_{ii}^{(k)}) + E_k, \quad (4.2)$$

где E_k — симметричная матрица внедиагональных элементов, и докажем сначала, что

$$\|E_k\|_E \rightarrow 0 \quad \text{при} \quad k \rightarrow \infty. \quad (4.3)$$

Из уравнений (3.3) имеем

$$\sum_{i \neq p, q} [(a_{ip}^{(k)})^2 + (a_{iq}^{(k)})^2] = \sum_{i \neq p, q} [(a_{ip}^{(k-1)})^2 + (a_{iq}^{(k-1)})^2], \quad (4.4)$$

и, следовательно, сумма квадратов внедиагональных элементов, за исключением элементов (p, q) и (q, p) , остается постоянной. Но эти два элемента становятся нулями, и, следовательно,

$$\|E_k\|_E^2 = \|E_{k-1}\|_E^2 - 2(a_{pq}^{(k-1)})^2, \quad (4.5)$$

а так как $a_{pq}^{(k-1)}$ есть максимальный по модулю элемент E_{k-1} , то

$$\|E_k\|_E^2 \leq [1 - 2/(n^2 - n)] \|E_{k-1}\|_E^2 \leq [1 - 2/(n^2 - n)]^k \|E_0\|_E^2. \quad (4.6)$$

Неравенство (4.6) является довольно грубой оценкой скорости сходимости. Если

$$N = \frac{1}{2} (n^2 - n), \quad (4.7)$$

то

$$\|E_{rN}\|_E^2 \leq (1 - 1/N)^{rN} \|E_0\|_E^2 < e^{-r} \|E_0\|_E^2. \quad (4.8)$$

Мы увидим, что это неравенство дает значительную недооценку скорости сходимости. Однако оно показывает, что

$$\|E_{rN}\|_E^2 < \varepsilon^2 \|E_0\|_E^2, \quad \text{когда} \quad r > 2 \ln(1/\varepsilon), \quad (4.9)$$

так что, например, для $\varepsilon = 2^{-t}$ это дает

$$r > 2 \ln 2^t \approx 1,39t. \quad (4.10)$$

Сходимость к фиксированной диагональной матрице

5. Мы должны еще показать, что A_k сходятся к фиксированной диагональной матрице. Предположим, что мы продолжали итерации до тех пор, пока не выполнилось неравенство

$$\|E_k\|_E < \varepsilon. \quad (5.1)$$

Тогда из (4.2) видим, что если собственные значения λ_i матрицы A_k , а, следовательно, и A_0 и $a_{ii}^{(k)}$ расположены в порядке возрастания, то соответствующие значения в двух последовательностях отличаются меньше чем на ε (гл. 2, § 44). Поэтому $a_{ii}^{(k)}$, расположенные в некотором порядке, лежат в интервалах длины 2ε с центрами в λ_i . Так как ε может быть произвольно малым, то мы можем локализовать собственные значения с любой нужной точностью.

Теперь покажем, что каждое $a_{ii}^{(k)}$ действительно сходится к определенному λ_i . Предположим сначала, что λ_i различны и определим ε соотношением

$$0 < 4\varepsilon = \min_{i \neq j} |\lambda_i - \lambda_j|. \quad (5.2)$$

Пусть k выбрано так, что (5.1) выполняется; согласно (4.6) оно выполняется для всех последующих E_r . При таком выборе ε интервалы с центрами

в λ_i , очевидно, разделяются, и поэтому в каждом интервале находится один элемент $a_{jj}^{(k)}$. Мы можем предположить, что λ_i занумерованы так, что каждое $a_{ii}^{(k)}$ лежит в λ_i -интервале. Покажем, что оно должно оставаться в этом интервале на всех последующих шагах.

Предположим, что следующее вращение осуществляется в плоскости (p, q) ; тогда элементы (p, p) и (q, q) будут единственными изменяющимися диагональными элементами. Следовательно, $a_{pp}^{(k+1)}$ и $a_{qq}^{(k+1)}$ должны быть двумя диагональными элементами из интервалов λ_p и λ_q . Покажем, что, например, в λ_p -интервале должно лежать только $a_{pp}^{(k+1)}$. Имеем

$$\begin{aligned} a_{qq}^{(k+1)} - \lambda_p &= a_{pp}^{(k)} \sin^2 \theta - 2a_{pq}^{(k)} \cos \theta \sin \theta + a_{qq}^{(k)} \cos^2 \theta - \lambda_p = \\ &= (a_{pp}^{(k)} - \lambda_p) \sin^2 \theta + (a_{qq}^{(k)} - \lambda_q + \lambda_q - \lambda_p) \cos^2 \theta - 2a_{pq}^{(k)} \cos \theta \sin \theta, \end{aligned} \quad (5.3)$$

откуда

$$\begin{aligned} a_{qq}^{(k+1)} - \lambda_p &\geq |\lambda_q - \lambda_p| \cos^2 \theta - |a_{pp}^{(k)} - \lambda_p| \sin^2 \theta - |a_{qq}^{(k)} - \lambda_q| \cos^2 \theta - \\ &- |a_{pq}^{(k)}| \geq 4\epsilon \cos^2 \theta - \epsilon \sin^2 \theta - \epsilon \cos^2 \theta - \epsilon = 2\epsilon \cos 2\theta. \end{aligned} \quad (5.4)$$

Однако

$$|\operatorname{tg} 2\theta| = |2a_{pq}^{(k)} / (a_{pp}^{(k)} - a_{qq}^{(k)})| \leq 2\epsilon / 2\epsilon = 1. \quad (5.5)$$

Следовательно, для θ в пределах (3.6) имеем $|2\theta| < \frac{1}{4}\pi$, и поэтому

$$|a_{qq}^{(k+1)} - \lambda_p| \geq 2^{1/2}\epsilon. \quad (5.6)$$

Это показывает, что $a_{qq}^{(k+1)}$ не находится в λ_p -интервале.

6. Введение кратных собственных значений не представляет большой трудности. Определим ϵ так, чтобы (5.2) выполнялось для всех различных λ_i и λ_j . Мы знаем, что если λ_i есть корень кратности s , то именно s коэффициентов из $a_{jj}^{(k)}$ лежат в λ_i -интервале (гл. 2, § 17). Доказательство того, что $a_{pp}^{(k)}$ и $a_{qq}^{(k)}$ не могут перейти в другие λ_i -интервалы, остается неизменным. Итак, во всех случаях, начиная с некоторого момента, каждый диагональный элемент матриц A_k остается в фиксированном λ_i -интервале, когда k возрастает, а длина интервала стремится к нулю, когда $k \rightarrow \infty$.

Циклический метод Якоби

7. На поиск наибольшего внедиагонального элемента тратится время работы вычислительной машины, и обычно элемент, подлежащий исключению, выбирается более простым способом. Возможно, что самой простой схемой является выбор элементов в последовательном порядке $(1, 2)$, $(1, 3), \dots, (1, n)$; $(2, 3)$, $(2, 4), \dots, (2, n)$; $\dots$; $(n-1, n)$, после чего снова возвращаемся к элементу $(1, 2)$. Эта схема иногда называется *частным* циклическим методом Якоби, в отличие от обобщения метода, в котором все внедиагональные элементы также исключаются по одному разу в некоторой последовательности из N вращений, но не обязательно в приведенном выше порядке. В циклическом методе (частном или общем) будем называть *циклом* любую последовательность из N вращений.

Хенричи (1958а) показал, что *частный* циклический метод Якоби действительно сходится, если углы вращения соответствующим образом ограничены, но доказательство очень сложное. В § 15 мы опишем вариант циклического метода, для которого очевидна квадратичная сходимость.

Круги Гершгорина

8. Предположим, что собственные значения λ_i матрицы A_0 различны, что

$$\min_{i \neq j} |\lambda_i - \lambda_j| = 2\delta \quad (8.1)$$

и что мы уже получили матрицу A_k , для которой

$$|a_{ii}^{(k)} - \lambda_i| < \frac{1}{4} \delta, \quad \max_{i \neq j} |a_{ij}| = \varepsilon, \quad (8.2)$$

где

$$(n-1)\varepsilon < \frac{1}{4} \delta. \quad (8.3)$$

В этом случае круги Гершгорина обязательно разделены, так как

$$|a_{ii}^{(k)} - a_{jj}^{(k)}| \geq |\lambda_i - \lambda_j| - |a_{ii}^{(k)} - \lambda_i| - |a_{jj}^{(k)} - \lambda_j| > \frac{3}{2} \delta. \quad (8.4)$$

Поэтому имеем

$$|a_{ii}^{(k)} - \lambda_i| < (n-1)\varepsilon. \quad (8.5)$$

Теперь, если умножим строку i на ε/δ и столбец i на δ/ε , то i -й круг Гершгорина будет содержаться в круге

$$|a_{ii}^{(k)} - \lambda| < (n-1)\varepsilon^2/\delta, \quad (8.6)$$

тогда как оставшиеся будут содержаться в кругах

$$|a_{jj}^{(k)} - \lambda| < (n-2)\varepsilon + \delta \quad (j \neq i). \quad (8.7)$$

Очевидно, что i -й круг изолирован от остальных, поэтому из (8.6) мы видим, что когда абсолютные значения внедиагональных элементов становятся меньше ε , то при достаточно малом ε диагональные элементы отличаются от собственных значений на величины порядка ε^2 . Заметим, что если A имеет несколько кратных собственных значений, то мы все же можем получить тот же результат для любого из простых собственных значений.

Асимптотическая квадратичная сходимость методов Якоби

9. Мы только что показали, что можно получить очень хорошие оценки для собственных значений, когда внедиагональные элементы стали малы. Это важно потому, что сходимость любого метода Якоби должна стать более быстрой на последующих шагах. Известно, что для различных вариантов метода Якоби сходимость в действительности становится асимптотически квадратичной, если собственные значения различны. Первый результат такого типа был получен Хенричи (1958а). Самыми точными результатами, полученными до сих пор, в обозначениях §§ 4 и 8 являются следующие.

(i) Для классического метода Якоби Шенхаге (1961) показал, что если

$$|\lambda_i - \lambda_j| \geq 2\delta \quad (i \neq j) \quad (9.1)$$

и мы достигли такого этапа, на котором

$$\|E_r\|_E < \frac{1}{2} \delta, \quad (9.2)$$

то

$$\|E_{r+N}\|_E \leq \left(\frac{1}{2}n - 1\right)^{1/2} \|E_r\|_E^2 / \delta. \quad (9.3)$$

Шенхаге показал далее, что если (9.1) выполняется за исключением p пар совпадающих собственных значений, то

$$\|E_{r+r'}\|_E \leq \frac{1}{2} n \|E_r\|_E^2 / \delta, \quad (9.4)$$

где

$$r' = N + p(n - 2). \quad (9.5)$$

(ii) Для общего циклического метода Якоби Уилкинсон (1962с) показал, что при условиях (9.1) и (9.2)

$$\|E_{r+N}\|_E \leq \frac{1}{2} [n(n - 1)]^{1/2} \|E_r\|_E^2 / \delta. \quad (9.6)$$

Шенхаге отметил, что тот же результат верен даже тогда, когда A имеет любое число пар совпадающих собственных значений при условии, что вращения, соответствующие этим парам, выполняются первыми.

(iii) Для частного циклического метода Якоби Уилкинсон (1962с) показал, что

$$\|E_{r+N}\|_E \leq \|E_r\|_E^2 / (2^{1/2} \delta). \quad (9.7)$$

Напомним, что доказательства этих результатов показывают лишь то, что если процессы сходятся, то они асимптотически сходятся квадратично. Никогда не было доказано, что общий циклический метод Якоби сходится (см. однако, § 15).

Ближние и кратные собственные значения

10. Все оценки скорости сходимости содержат множитель $1/8$, и это могло бы навести на мысль, что когда δ мало, начало квадратичной сходимости будет задерживаться. Если δ равно нулю, до доказательства не проходят, хотя было показано, что когда нет собственных значений кратности больше чем два, то квадратичная сходимость может быть гарантирована, если порядок исключения элементов несколько изменен (см., например, Шенхаге (1961)). В исследовании скорости сходимости, когда существуют собственные значения кратности больше чем два, сделано пока немного. Наш опыт работы с частным циклическим методом Якоби привел к мысли, что на практике очень близкие и кратные собственные значения обычно влияют на число итераций, необходимых для приведения внедиагональных элементов ниже предписанного уровня, существенно меньше, чем можно было бы ожидать, учитывая настоящий уровень наших знаний. Для матриц до 50-го порядка и для длин слов от 32 до 48 двоичных разрядов общее число циклов в среднем равно 5—6, а число итераций, необходимое для матриц, имеющих кратные собственные значения, часто оказывалось несколько меньше чем среднее. Следующие рассуждения показывают, как это может происходить.

Во-первых, если все собственные значения равны λ_1 , то матрица есть $\lambda_1 I$ и не нужно ни одной итерации. Предположим, что взят менее крайний случай, когда все собственные значения очень близки, так что

$$\lambda_i = a + \delta_i, \quad (10.1)$$

где δ_i мало по сравнению с a . Тогда

$$A = R^T \text{diag}(a + \delta_i) R \quad (10.2)$$

для некоторой ортогональной матрицы R и, следовательно,

$$A = aI + R^T \text{diag}(\delta_i) R. \quad (10.3)$$

Далее мы имеем

$$\|R^T \text{diag}(\delta_i) R\|_E = (\sum \delta_i^2)^{1/2}, \quad (10.4)$$

что показывает, что внедиагональные элементы должны быть малы с самого начала. Вычисления с A идентичны вычислениям с $(A - aI)$, а эта матрица уже не является специальной в первоначальном смысле. Но так как все элементы $(A - aI)$ малы, то мы можем ожидать, что для приведения внедиагональных элементов матрицы A ниже предписанного уровня потребуется меньше времени, чем для общей симметричной матрицы. Действительно, если $\max |\delta_i| = O(10^{-r})$, то мы могли бы ожидать, что приведем внедиагональные элементы A ниже уровня 10^{-t-r} за то же самое время, что в обычном случае до уровня 10^{-t} .

Например, для симметричной матрицы B

$$B = \begin{bmatrix} 0,2512 & 0,0014 & 0,0017 \\ & 0,2511 & 0,0026 \\ & & 0,2507 \end{bmatrix} \quad (10.5)$$

(мы даем лишь верхний треугольник элементов) углы вращения, очевидно, такие же, как для матрицы C ,

$$C = 100B - 25I = \begin{bmatrix} 0,1200 & 0,1400 & 0,1700 \\ & 0,1100 & 0,2600 \\ & & 0,0700 \end{bmatrix} \quad (10.6)$$

Однако если мы имеем дело с матрицей C , то должны помнить, что нам нужно привести ее внедиагональные элементы лишь до уровня 10^{-2} для того, чтобы сделать внедиагональные элементы матрицы B порядка 10^{-4} .

Далее можно показать, что если матрица A имеет собственное значение a кратности $(n - 1)$, то она будет приведена к диагональному виду меньше чем за один цикл. Обозначая другое собственное значение через b , A можно представить в виде

$$A = R^T \begin{bmatrix} a & & & \\ & a & & \\ & & \ddots & \\ & & & a \end{bmatrix} R = aI + R^T \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & b - a \end{bmatrix} R. \quad (10.7)$$

Матрица aI не влияет на процесс Якоби, а элемент (i, j) второй матрицы в правой части (10.7) равен $(b - a) r_{ni} r_{nj}$. Отсюда сразу же заключаем, что требуется $(n - 1)$ вращений для того, чтобы исключить все элементы первой строки матрицы A , а в действительности для того, чтобы исключить все внедиагональные элементы.

Численные примеры

11. Табл. 1 и 2 иллюстрируют как сам процесс, так и квадратичную сходимость на последних этапах. Заметим, что нам не нужно делать преобразования и строки и столбца, за исключением четырех элементов,

Таблица 1

$$A_0 = \begin{bmatrix} 0,6532 & 0,2165 & 0,0031 \\ & 0,4105 & 0,0052 \\ & & 0,2132 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (1,2)} \\ \operatorname{tg} 2\theta = 2 \times 0,2165 / (0,6532 - 0,4105) \\ \cos \theta = 0,8628, \sin \theta = 0,5055 \end{array}$$

$$A_1 = \begin{bmatrix} 0,7800 & 0,0000 & 0,0053 \\ & 0,2836 & 0,0029 \\ & & 0,2132 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (1,3)} \\ \operatorname{tg} 2\theta = 2 \times 0,0053 / (0,7800 - 0,2132) \\ \cos \theta = 1,0000, \sin \theta = 0,0094 \end{array}$$

Заметим, что так как $a_{13}^{(1)}$ мал, то мал θ .

$$A_2 = \begin{bmatrix} 0,7801 & 0,0000 & 0,0000 \\ & 0,2836 & 0,0029 \\ & & 0,2132 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (2,3)} \\ \operatorname{tg} 2\theta = 2 \times 0,0029 / (0,2836 - 0,2132) \\ \cos \theta = 0,9991, \sin \theta = 0,0411 \end{array}$$

Элемент (1,2) остался нулем с рабочей точностью. Аналогично элемент (2,3) остался мал; в действительности он не меняется с рабочей точностью.

$$A_3 = \begin{bmatrix} 0,7801 & 0,0000 & 0,0000 \\ & 0,2837 & 0,0000 \\ & & 0,2131 \end{bmatrix}$$

Матрица диагонализирована с рабочей точностью.

Таблица 2

$$A_0 = \begin{bmatrix} 0,2500 & 0,0012 & 0,0014 \\ & 0,2500 & 0,0037 \\ & & 0,6123 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (1,2)} \\ \operatorname{tg} 2\theta = 2 \times 0,0012 / (0,2500 - 0,2500) \\ \cos \theta = \sin \theta = 0,7071 \end{array}$$

$$A_1 = \begin{bmatrix} 0,2512 & 0,0000 & 0,0036 \\ & 0,2488 & 0,0016 \\ & & 0,6123 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (1,3)} \\ \operatorname{tg} 2\theta = 2 \times 0,0036 / (0,2512 - 0,6123) \\ \cos \theta = 1,0000, \sin \theta = -0,0099 \end{array}$$

Следующие два вращения с малыми углами и близость элементов (1,1) и (2,2) не имеет значения.

$$A_2 = \begin{bmatrix} 0,2512 & 0,0000 & 0,0000 \\ & 0,2488 & 0,0016 \\ & & 0,6124 \end{bmatrix} \begin{array}{l} \text{Вращение в плоскости (2,3)} \\ \operatorname{tg} 2\theta = 2 \times 0,0016 / (0,2488 - 0,6124) \\ \cos \theta = 1,0000, \sin \theta = -0,0044 \end{array}$$

$$A_3 = \begin{bmatrix} 0,2512 & 0,0000 & 0,0000 \\ & 0,2488 & 0,0000 \\ & & 0,6124 \end{bmatrix}$$

стоящих на пересечении соответствующих строк и столбцов, причем два из этих элементов становятся нулями. Храня и вычисляя лишь наддиагональные элементы и предполагая, что соответствующие поддиагональные элементы должны быть равны им, мы сохраняем точную симметрию. Поэтому требуется выполнить примерно $4n$ умножений при каждом вращении и $2n^3$ умножений при каждом цикле. Элементы, которые изменяются в матрице шестого порядка при вращении в плоскости (3, 5), указаны на

следующей схеме:

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ & & \vdots & & \vdots & \\ & \times & \times & \times & \times & \times \\ & & \vdots & & \vdots & \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \vdots & \\ & & & & \times & \times \\ & & & & & \times \\ & & & & & \times \end{bmatrix}. \quad (11.1)$$

Пример в табл. 1 выбран так, чтобы после первого вращения все внедиагональные элементы были малыми по сравнению с разностями между диагональными элементами. Все последующие вращения имеют малые углы и внедиагональные элементы, которые в данный момент не делаются нулями, изменяются так мало, что с рабочей точностью в четыре десятичных знака остаются постоянными.

Пример в табл. 2 выбран для иллюстрации того факта, что наличие пары близких диагональных элементов не задерживает начало квадратичной сходимости. Мы взяли первые два диагональных элемента равными, а все внедиагональные элементы малыми. Первое вращение (в плоскости (1, 2)) имеет угол, равный $\frac{1}{4}\pi$. Оно делает элемент (1, 2) нулем и оставляет другие внедиагональные элементы величинами прежнего порядка. Угол второго вращения (в плоскости (1, 3)) мал, так как первый и третий диагональные элементы не близки. То же самое относится к вращению (2, 3), и в конце цикла все внедиагональные элементы становятся нулями с рабочей точностью.

Очевидно, что если есть любое число пар близких (или совпадающих) диагональных элементов, то применимо такое же рассуждение, но оно неверно, если существует группа из трех и более элементов.

Вычисление $\cos \theta$ и $\sin \theta$

12. До сих пор вопрос о вычислении угла вращения не затрагивался, но следует отметить, что оно не покрывается анализом, данным в гл. 3, § 20. На практике требуется значительное внимание к деталям, если мы хотим обеспечить, чтобы:

(i) $\cos^2 \theta + \sin^2 \theta$ было очень близко к единице, так что матрица преобразования действительно была бы почти ортогональна и преобразованная матрица почти подобна исходной;

(ii) $(a_{qq}^{(k)} - a_{pp}^{(k)}) \cos \theta \sin \theta + a_{pq}^{(k)} (\cos^2 \theta - \sin^2 \theta)$ было очень мало, так что мы вправе заменить $a_{pq}^{(k+1)}$ нулем.

Опустим индекс k и напомним

$$\operatorname{tg} 2\theta = 2a_{pq}/(a_{pp} - a_{qq}) = x/y, \quad (12.1)$$

где

$$x = \pm 2a_{pq}, \quad y = |a_{pp} - a_{qq}|. \quad (12.2)$$

Тогда для θ , лежащего в пределах $|\theta| < \pi/4$, формально имеем

$$\begin{cases} 2 \cos^2 \theta = [(x^2 + y^2)^{1/2} + y]/(x^2 + y^2)^{1/2}, \\ 2 \sin^2 \theta = [(x^2 + y^2)^{1/2} - y]/(x^2 + y^2)^{1/2}, \end{cases} \quad (12.3)$$

где $\cos \theta$ должен выбираться положительным, а $\sin \theta$ иметь такой же знак, как и $\operatorname{tg} 2\theta$.

Теперь, если x мало по сравнению с y , уравнения (12.3) дают очень неудовлетворительное определение $\sin \theta$. Для десятизначных десятичных вычислений как с фиксированной, так и с плавающей запятой (12.3) дает $\sin \theta = 0$ и $\cos \theta = 1$ для всех значений x и y , удовлетворяющих неравенству $|x| < 10^{-5} |y|$. Итак, хотя условие (i) выполняется, условие (ii) — нет.

С другой стороны, если используется плавающая запятая, определение $\cos \theta$ из (12.3) вполне удовлетворительно. Действительно, из гл. 3, §§ 6 и 10 имеем

$$\left. \begin{aligned} u &= fl(x^2 + y^2) \equiv (x^2 + y^2)(1 + \varepsilon_1) & (|\varepsilon_1| < (2,00002) 2^{-t}), \\ v &= fl(u^{1/2}) \equiv u^{1/2}(1 + \varepsilon_2) & (|\varepsilon_2| < (1,00001) 2^{-t}), \\ w &= fl(v + y) \equiv (v + y)(1 + \varepsilon_3) & (|\varepsilon_3| \leq 2^{-t}), \\ z &= fl(w/2v) \equiv w(1 + \varepsilon_4)/2v & (|\varepsilon_4| \leq 2^{-t}), \\ c &= fl(z^{1/2}) \equiv z^{1/2}(1 + \varepsilon_5) & (|\varepsilon_5| < (1,00001) 2^{-t}). \end{aligned} \right\} \quad (12.4)$$

Объединяя эти результаты, получаем

$$c \equiv \cos \theta (1 + \varepsilon_6) \quad (|\varepsilon_6| < (3,0003) 2^{-t}), \quad (12.5)$$

где $\cos \theta$ есть точное значение, соответствующее x и y . Здесь и далее оценки, которые мы даем, неточные; часто их можно будет улучшить с помощью более подробного анализа. Значение s для $\sin \theta$ может быть получено из соотношения

$$2 \sin \theta \cos \theta = x/(x^2 + y^2)^{1/2}. \quad (12.6)$$

Используя плавающую арифметику и вычисленное значение c , имеем

$$s = fl(x/2cv) \equiv x(1 + \varepsilon_7)/2cv \quad (|\varepsilon_7| < (2,00002) 2^{-t}), \quad (12.7)$$

откуда, учитывая (12.4) и (12.5), можем заключить, что

$$s \equiv \sin \theta (1 + \varepsilon_8) \quad (|\varepsilon_8| < (6,0006) 2^{-t}). \quad (12.8)$$

Итак, и s , и c имеют малые относительные ошибки по сравнению с точными значениями, соответствующими данным x и y . Теперь ясно, что условие (ii) при вычисленных значениях выполняется.

13. Если могут быть накоплены скалярные произведения, то для того, чтобы получить аналогичные результаты с фиксированной запятой, очень хорош следующий способ. Вычисляем точно $x^2 + y^2$ и затем выбираем наибольшее целое число k , для которого

$$2^{2k}(x^2 + y^2) < 1. \quad (13.1)$$

Далее выполняются вычисления, использующие $2^k x$ и $2^k y$ вместо x и y в уравнении (12.3) для $\cos^2 \theta$ и в уравнении (12.6) для $\sin \theta$.

Если же скалярные произведения не могут быть накоплены, то мы можем написать

$$2 \cos^2 \theta = \frac{[1 + (y/x)^2]^{1/2} + y/x}{[1 + (y/x)^2]^{1/2}}, \quad (13.2)$$

$$2 \cos \theta \sin \theta = 1/[1 + (y/x)^2]^{1/2} \quad (13.3)$$

для $|x| \geq y$ и соответствующие выражения для $y > |x|$ (ср. гл. 3, § 30).

Более простое определение углов вращения

14. На некоторых вычислительных машинах определение тригонометрических функций по методам §§ 12 и 13 предъявляет более тяжелые требования ко времени и памяти, чем это желательно, поэтому были разработаны другие более простые методы определения углов вращения.

Так как ортогональное преобразование оставляет собственные значения неизменными, то нам не обязательно определять угол вращения так, чтобы сделать соответствующий элемент нулем. При условии, что мы существенно понижаем его значение, можно ожидать сходимость, так как $\|E_k\|_E^2$ уменьшается на $2(a_{pq}^{(k)})^2 - 2(a_{pq}^{(k+1)})^2$. Однако для того, чтобы сохранить на последних этапах квадратичную сходимость, важно, чтобы углы вращений стремились к углам, вычисленным по общепринятому методу, когда внедиагональные элементы становятся малыми.

В одном из таких методов угол вращения ϕ определяется соотношением

$$T = \operatorname{tg} \frac{1}{2} \phi = a_{pq}/2(a_{pp} - a_{qq}) \quad (14.1)$$

для ϕ , лежащего в пределах $(-\pi/4, \pi/4)$ и $\phi = \pm \pi/4$ в противном случае, причем знак берется равным знаку правой части (14.1). Сравнивая (14.1) и (12.1), мы видим, что

$$\operatorname{tg} 2\theta = 4 \operatorname{tg} \frac{1}{2} \phi \quad (14.2)$$

и, следовательно,

$$\phi - \theta \sim \frac{5}{4} \theta^3 \quad \text{при} \quad \theta \rightarrow 0. \quad (14.3)$$

Мы можем вычислить $\sin \phi$ и $\cos \phi$ по формулам

$$\sin \phi = 2T/(1 + T^2), \quad \cos \phi = (1 - T^2)/(1 + T^2), \quad (14.4)$$

и, так как $|T| < 1$, вычисление очень просто как в фиксированной, так и в плавающей запятой. Легко проверить, что при таком выборе угла влияние вращения эквивалентно замене $a_{pq}^{(k)}$ на $\alpha a_{pq}^{(k)}$, где α удовлетворяет неравенству

$$0 \leq \alpha \leq \frac{1}{4} (1 + 2^{1/2}) \quad (14.5)$$

и близко к нулю для большинства углов, определенных из (14.1). Оказалось, что скорость сходимости с таким определением углов вращения сравнима со скоростью сходимости, соответствующей общепринятому выбору.

Циклический метод Якоби с преградами

15. Максимальное снижение значения $\|E_k\|_E^2$, которое может быть достигнуто при вращении в плоскости (p, q) , равно $2(a_{pq}^{(k)})^2$. Если $a_{pq}^{(k)}$ много меньше, чем общий уровень внедиагональных элементов, то вращение (p, q) приносит мало пользы. Это привело к созданию варианта метода, известного как *циклический метод Якоби с преградами*.

С каждым циклом мы свяжем значение преграды и при выполнении преобразований этого цикла опустим те вращения, которые основаны на внедиагональных элементах, меньших значения преграды. На DEUCE при работе с фиксированной запятой использовалась совокупность преград 2^{-3} , 2^{-6} , 2^{-10} , 2^{-18} , 2^{-30} . Все последующие циклы имели значение преграды, равное 2^{-30} , и так как это есть наименьшее допустимое число на DEUCE, то на пятом и последующих циклах пропускаются только нули. Процесс заканчивается, когда пропускается $1/2(n^2 - n)$ последовательных элементов. Опыт работы на DEUCE позволяет высказать предположение, что эта модификация дает умеренное улучшение в скорости. Отметим, что сходимость этого варианта метода гарантируется, так как может существовать лишь конечное число итераций, соответствующих любой заданной преграде.

На SWAC использовалась совокупность преград, заданная последовательностью $k, k^2, k^3, \dots$ для некоторого заранее предписанного значения k (Попе и Томпкинс (1957)). Итерации с каждой преградой продолжались до тех пор, пока все внедиагональные элементы не становились меньше ее.

Вычисление собственных векторов

16. Если последнее вращение обозначить через R_s , то мы имеем

$$(R_s \dots R_2 R_1) A_0 (R_1^T R_2^T \dots R_s^T) = \text{diag}(\lambda_i) \quad (16.1)$$

с рабочей точностью. Поэтому собственные векторы матрицы A_0 являются столбцами матрицы P_s , определенной равенством

$$P_s = R_1^T R_2^T \dots R_s^T. \quad (16.2)$$

Если требуются собственные векторы, то на каждом этапе процесса мы можем сохранять текущее произведение P_k , где

$$P_k = R_1^T R_2^T \dots R_k^T, \quad (16.3)$$

для чего нужны n^2 ячеек памяти дополнительно к $1/2(n^2 + n)$ ячейкам, необходимым для хранения наддиагональных элементов последовательных матриц A_k . Если мы принимаем эту схему, то автоматически получаем все n собственных векторов.

Если требуется лишь несколько собственных векторов, то можно сэкономить на времени вычислений, если мы будем запоминать $\cos \theta$ и $\sin \theta$ в плоскости каждого вращения. Предположим, что требуется только один собственный вектор, который соответствует j -му диагональному элементу конечной матрицы $\text{diag}(\lambda_i)$, тогда собственный вектор v_j определяется так:

$$v_j = R_1^T R_2^T \dots R_s^T e_j. \quad (16.4)$$

Если умножим e_j слева последовательно на $R_s^T, R_{s-1}^T, \dots, R_1^T$, то можем получить j -й собственный вектор, не вычисляя остальных. Для вычисления P_s необходимо выполнить в n раз больше умножений, чем для вычисления v_j , поэтому для вычисления r векторов число операций в r/n раз меньше, чем для вычисления P_s .

Так как в среднем требуется пять или шесть циклов, то для того, чтобы запоминать вращения, нужна память в несколько раз больше, чем для запоминания P_s , но на вычислительной машине с памятью магнитного типа это, по-видимому, не ограничивает порядок матриц, с которыми можно будет иметь дело.

Численный пример

17. Способ вычисления отдельного вектора иллюстрируется в табл. 3. Здесь мы вычисляем собственный вектор, соответствующий второму собственному значению примера из табл. 1. Последний вектор имеет евклидо-

Таблица 3

$$v_2 = \begin{bmatrix} 0,8628 & -0,5055 & 0 \\ 0,5055 & 0,8628 & 0 \\ 0 & 0 & 1,0000 \end{bmatrix} \begin{bmatrix} 1,000 & 0 & -0,0094 \\ 0 & 1,0000 & 0 \\ 0,0094 & 0 & 1,0000 \end{bmatrix} \times \\ \times \begin{bmatrix} 1,0000 & 0 & 0 \\ 0 & 0,9991 & -0,0411 \\ 0 & 0,0411 & 0,9991 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Умножение на последовательные матрицы дает

$$\begin{bmatrix} 0,0000 \\ 0,9991 \\ 0,0411 \end{bmatrix}, \begin{bmatrix} -0,0004 \\ 0,9991 \\ 0,0411 \end{bmatrix} \text{ и окончательно } \begin{bmatrix} -0,5054 \\ 0,8618 \\ 0,0411 \end{bmatrix}$$

ву норму, равную единице, так как P_s ортогональна. Это может быть использовано для проверки выполненных вычислений и как мера накопления ошибок округления.

Анализ ошибок метода Якоби

18. Метод Якоби не вполне охватывается общим анализом, данным в гл. 3, §§ 20—36, так как вычисление $\cos \theta$ и $\sin \theta$ несколько отличается от того, что приведено там. Однако мы показали, что вычисленные $\cos \theta$ и $\sin \theta$ имеют малые относительные ошибки по сравнению с точными значениями, соответствующими текущей вычисленной матрице, поэтому изменения в анализе главы 3 очень просты.

Предлагаем в качестве упражнения показать, что в главе 3 в уравнении (21.5) множитель $6 \cdot 2^{-t}$ можно заменить на $9 \cdot 2^{-t}$, если $\sin \theta$ и $\cos \theta$ вычисляются согласно § 12. Если выполняется r полных циклов и окончательно вычисленная диагональная матрица обозначается через $\text{diag}(\mu_i)$, то соответственно уравнению (27.2) главы 3 имеем

$$\|\text{diag}(\mu_i) - RA_0R^T\|_E \leq 18 \cdot 2^{-t} n^{3/2} (1 + 9 \cdot 2^{-t})^{r(4n-7)} \|A_0\|_E, \quad (18.1)$$

где R есть точная ортогональная матрица, полученная как точное произведение точных плоских вращений, соответствующих вычисленным A_k . Если возьмем r равным шести, то

$$\|\text{diag}(\mu_i) - RA_0R^T\|_E \leq 108 \cdot 2^{-t} n^{3/2} (1 + 9 \cdot 2^{-t})^{6(4n-7)} \|A_0\|_E \quad (18.2)$$

Это означает, что если λ_i являются точными собственными значениями матрицы A_0 и как λ_i , так и μ_i расположены в возрастающем порядке, то

$$\left[\frac{\sum (\mu_i - \lambda_i)^2}{\sum \lambda_i^2} \right]^{1/2} \leq 108 \cdot 2^{-t} n^{3/2} (1 + 9 \cdot 2^{-t})^{6(4n-7)}. \quad (18.3)$$

Как обычно, последний множитель на практике не имеет большого значения, так как если правая часть должна быть достаточно малой, то этот множитель почти равен единице. Правая часть (18.3) есть максимальная верхняя оценка, и мы могли бы ожидать, что статистическое распределение ошибок округления обеспечит оценку такого порядка:

$$\left[\frac{\sum (\mu_i - \lambda_i)^2}{\sum \lambda_i^2} \right]^{1/2} \leq 11 \cdot 2^{-t} n^{3/4}.$$

Если мы возьмем, например, $t = 40$ и $n = 10^4$, то правая часть равна примерно 10^{-8} . Это показывает, что, вероятно, на практике время или объем памяти должны быть более ограничивающими факторами, чем точность вычисленных собственных значений.

Точность вычисленных собственных векторов

19. Мы были в состоянии получить удовлетворительные строгие *априорные* оценки для ошибок в собственных значениях, но не можем надеяться получить аналогичные оценки для собственных векторов. В общем случае их точность неизбежно будет зависеть от разделения собственных значений. Если мы напомним

$$\text{diag}(\mu_i) - RA_0R^T = F, \quad (19.1)$$

то

$$\text{diag}(\mu_i) = R(A_0 + G)R^T, \quad (19.2)$$

где

$$G = R^TFR, \quad \|G\|_E = \|F\|_E. \quad (19.3)$$

Даже если бы мы могли вычислить R^T точно, то она дала бы нам лишь собственные векторы матрицы $(A_0 + G)$.

В действительности мы не можем получить точную матрицу R^T . Мы будем иметь лишь вычисленные вращения и будем делать ошибки при их перемножении. Если обозначить вычисленное произведение через $\bar{R}^T$, то, как в гл. 3, § 25, с множителем $9 \cdot 2^{-t}$ для x , имеем

$$\|\bar{R}^T - R^T\|_E \leq 9 \cdot 2^{-t} m^2 (1 + 9 \cdot 2^{-t})^{r(2n-4)}, \quad (19.4)$$

и снова, беря $r = 6$ и предполагая разумное распределение ошибок, мы могли бы ожидать приблизительно такую оценку:

$$\|\bar{R}^T - R^T\|_E \leq 8n2^{-t}. \quad (19.5)$$

Поэтому вычисленная матрица $\bar{R}^T$ будет очень близка к точно ортогональной матрице R^T .

Итак, мы получили оценку для отклонения $\bar{R}^T$ от R^T , точной матрицы собственных векторов матрицы $(A_0 + G)$ и уже имеем оценку для G . Оценки (19.4) и (19.5) не зависят от разделения собственных значений.

Даже когда матрица A_0 имеет несколько патологически близких собственных значений, соответствующие собственные векторы почти ортогональны, хотя они могут быть очень неточными; очевидно, что их линейная оболочка образует правильное подпространство и дает полную цифровую информацию. Возможно, что это самое хорошее свойство метода Якоби. Другие методы, которые мы опишем в этой главе, лучше метода Якоби по скорости и точности, но довольно утомительно получать ортогональные собственные векторы, соответствующие патологически близким или совпадающим собственным значениям.

Оценки ошибок для вычислений с фиксированной запятой

20. Методом, аналогичным описанному в гл. 3, § 29, мы можем вывести оценки ошибок в собственных значениях, полученных при использовании вычислений с фиксированной запятой. Точные оценки зависят от частного способа определения $\cos \theta$ и $\sin \theta$, но все они вида

$$\|\text{diag}(\mu_i) - RA_0R^T\|_2 \leq K_1 r n^{5/2} 2^{-t} \quad (20.1)$$

или

$$\|\text{diag}(\mu_i) - RA_0R^T\|_E \leq K_2 r n^{5/2} 2^{-t}, \quad (20.2)$$

в зависимости от того, пользуемся ли мы евклидовой или 2-нормой. Здесь опять r есть число циклов. Как и в гл. 3, §§ 32, 33, неравенство (20.1) получается в предположении, что A_0 первоначально нормирована так, что

$$\|A_0\|_2 \leq 1 - K_1 r n^{5/2} 2^{-t}. \quad (20.3)$$

Эту трудность требуется преодолевать, так как мы обычно не имеем надежной оценки для $\|A_0\|_2$ до тех пор, пока не выполним процесс. С другой стороны, (20.2) выводится в предположении, что A_0 первоначально нормирована так, что

$$\|A_0\|_E \leq 1 - K_2 r n^{5/2} 2^{-t}. \quad (20.4)$$

Эта трудность преодолевается легко, так как $\|A_0\|_E$ может быть вычислена очень просто. Такое нормирование гарантирует, что ни один элемент любой A_k не даст переполнения.

Снова по методу, аналогичному тому, который описан в гл. 3, § 34, находим, что $\bar{R}^T$, вычисленное произведение вычисленных вращений, удовлетворяет неравенству

$$\|\bar{R}^T - R^T\|_2 \leq K_3 r n^{5/2} 2^{-t}, \quad (20.5)$$

где R^T — точная матрица собственных векторов матрицы $(A_0 + G)$, и мы имеем или

$$\|G\|_2 \leq K_1 r n^{5/2} 2^{-t} \quad \text{или} \quad \|G\|_E \leq K_2 r n^{5/2} 2^{-t}. \quad (20.6)$$

Как и при вычислениях с плавающей запятой, $\bar{R}^T$ будет почти ортогональна для всех приемлемых значений n , независимо от того, насколько ее столбцы могут быть близки к собственным векторам матрицы A_0 .

Организационные проблемы

21. Неудобной особенностью метода Якоби является то, что при каждом преобразовании требуется доступ к строкам и столбцам текущей матрицы. Если матрица может быть размещена в быстродействующей памяти, то это не создает трудностей, но если быстродействующая

память недостаточна для этой цели, то такой путь организации метода Якоби кажется не очень удовлетворительным.

Чартрес (1962) описал *модифицированную* форму метода Якоби, которая хорошо приспособлена для использования вычислительной машины с внешней памятью магнитного типа. За детальным рассмотрением этого метода мы отсылаем читателя к статье Чартреса.

Метод Гивенса

22. Метод Якоби по существу носит итерационный характер. Следующие два метода, которые мы опишем, предназначены для приведения исходной матрицы к более простому виду и дают с помощью ортогональных подобных преобразований симметричную трехдиагональную матрицу (гл. 3, § 13). Такое приведение может быть выполнено неитерационными методами и, следовательно, оно требует намного меньше вычислений. Так как собственные значения симметричной трехдиагональной матрицы определяются сравнительно просто, эти методы весьма эффективны.

Первый такой метод, разработанный Гивенсом (1954), также основан на использовании плоских вращений. Он состоит из $(n - 2)$ основных шагов, на r -м из которых получаются нули в r -й строке и r -м столбце, при этом не портятся нули, полученные на предыдущих $(r - 1)$ шагах. В начале r -го основного шага первые $(r - 1)$ строк и столбцов образуют трехдиагональную матрицу. Для случая $n = 6$, $r = 3$ это иллюстрируется такой схемой:

$$\left[\begin{array}{cc|cccc} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 \\ \hline 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \end{array} \right] \begin{array}{l} \leftarrow \\ \leftarrow \\ \leftarrow \\ \leftarrow \\ \leftarrow \end{array} \quad (22.1)$$

↑ ↑ ↑

r -й основной шаг состоит из $(n - r - 1)$ вспомогательных шагов, на которых нули получаются последовательно в позициях $r + 2, r + 3, \dots, n$ r -й строки и r -го столбца; нуль в i -й позиции получается вращением в плоскости $(r + 1, i)$. В (22.1) мы подчеркнули элементы, которые исключаются на r -м основном шаге, и обозначили стрелками те строки и столбцы, которые затрагиваются. Сразу же видно, что первые $(r - 1)$ строк и столбцов целиком не меняются; хотя некоторые нулевые элементы участвуют в преобразовании, они заменяются линейной комбинацией нулей. Поэтому в действительности на r -м основном шаге мы имеем дело лишь с матрицей порядка $(n - r + 1)$ в нижнем правом углу текущей матрицы.

Для того чтобы облегчить сравнение с триангуляризацией плоскими вращениями, дадим сначала алгоритмическое описание процесса, предполагая, что на всех этапах сохраняется полная матрица. Текущие значения $\cos \theta$ и $\sin \theta$ помещаются на место тех элементов, которые в данный момент исключаются. Теперь r -й основной шаг состоит в следующем.

Для каждого значения i от $r + 2$ до n выполняем шаги от (i) до (iv):

(i) Вычисляем $x = (a_{r,r+1}^2 + a_{ri}^2)^{1/2}$.

(ii) Вычисляем $\cos \theta = a_{r,r+1}/x$, $\sin \theta = a_{ri}/x$. (Если $x = 0$, то берем $\cos \theta = 1$, $\sin \theta = 0$ и можем опустить (iii) и (iv).) Записываем $\cos \theta$ на

место a_{ri} , $\sin \theta$ на место a_{ir} и x на место $a_{r,r+1}$ и $a_{r+1,r}$, причем x в этих позициях есть правильное новое значение элементов.

(iii) Для каждого значения j от $r+1$ до n вычисляем $a_{j,r+1} \cos \theta + a_{ji} \sin \theta$ и $-a_{j,r+1} \sin \theta + a_{ji} \cos \theta$ и записываем их соответственно на место $a_{j,r+1}$ и a_{ji} .

(iv) Для каждого значения j от $r+1$ до n вычисляем $a_{r+1,j} \cos \theta + a_{ij} \sin \theta$ и $-a_{r+1,j} \sin \theta + a_{ij} \cos \theta$ и записываем их соответственно на место $a_{r+1,j}$ и a_{ij} .

На практике нам не нужно целиком выполнять шаги (iii) и (iv), так как вследствие симметрии (iv) в основном повторяет для строк те вычисления, которые в (iii) были выполнены для столбцов. Обычно работают лишь с наддиагональными элементами, и те из них, которые участвуют при вращениях, показаны в (11.1). Однако заметим, что если мы хотим сохранить значения $\cos \theta$ и $\sin \theta$, то нам нужны все n^2 ячеек памяти.

23. С учетом симметрии число умножений на r -м шаге приблизительно равно $4(n-r)^2$, следовательно, для полного приведения к трехдиагональной форме требуется выполнить около $\frac{4}{3}n^3$ умножений. Мы можем сравнить это с $2n^3$ умножениями на одном цикле процесса Якоби. Кроме того, есть примерно $\frac{1}{2}n^2$ извлечений квадратных корней как в преобразовании Гивенса, так и в одном цикле процесса Якоби. Вычисление углов вращения в методе Гивенса существенно проще. Если предположим, что в процессе Якоби выполняется шесть циклов, то для больших n преобразования Якоби требуют примерно в девять раз больше вычислений, чем преобразование Гивенса. При условии, что мы можем найти собственные значения трехдиагональной матрицы сравнительно быстро, процесс Гивенса будет работать намного быстрее процесса Якоби.

Процесс Гивенса на вычислительной машине с двухступенчатой памятью

24. Если мы представляем себе процесс Гивенса в том плане, как он использовался в § 22, то он сравнительно неэффективен для вычислительных машин с двухступенчатой памятью. Даже если мы воспользуемся симметрией, то найдем, что в каждый вспомогательный шаг включаются переносы как строки, так и столбца. Если, например, наддиагональные элементы записаны во внешней памяти по строкам, то на r -м основном шаге строки от r -й до n -й должны переноситься из внешней памяти и обратно на каждом вспомогательном шаге, хотя в каждой строке, кроме двух, на которых основано текущее вращение, изменяются лишь два элемента. Опишем сейчас модифицированный процесс, предложенный Роллетом и Уилкинсоном (1961) и Иогансеном (1961), в котором перенос каждой из строк от r -й до n -й в вычислительное запоминающее устройство и обратно делается только один раз за весь r -й основной шаг.

Первые $(r-1)$ строк и столбцов не играют никакой роли на r -м основном шаге, который формально аналогичен первому, но работает с матрицей порядка $(n-r+1)$ в нижнем правом углу. Таким образом, нет никакой потери общности в описании первого основного шага, и мы сделаем это ради простоты. *В быстросействующей памяти нам нужны четыре группы из n рабочих ячеек.* Первый основной шаг состоит из пяти больших этапов. (Так как элементы записаны частично во внешней памяти

и частично в быстродействующей, мы несколько отклонимся от алгоритмического языка, использованного ранее.)

(1) Переносим первую строку A в первую группу рабочих ячеек.

(2) Для каждого значения j от 3 до n вычисляем

$$x_j = [(a_{1,2}^{(j-1)})^2 + a_{1,j}^2]^{1/2}, \quad \cos \theta_{2,j} = a_{1,2}^{(j-1)}/x_j, \quad \sin \theta_{2,j} = a_{1,j}/x_j,$$

где

$$a_{1,2}^{(2)} = a_{1,2} \quad \text{и} \quad a_{1,2}^{(j)} = x_j.$$

Записываем $\cos \theta_{2,j}$ на место $a_{1,j}$ и запоминаем $\sin \theta_{2,j}$ во второй группе рабочих ячеек.

(3) Переносим вторую строку в третью группу рабочих ячеек. (У этой и всех последующих строк используются элементы, лежащие на диагонали и выше ее.)

Для каждого значения k от 3 до n выполняем шаги (4) и (5):

(4а) Переносим k -ю строку в четвертую группу рабочих ячеек.

(4б) Выполняем строчные и столбцовые операции, включающие $\cos \theta_{2,k}$ и $\sin \theta_{2,k}$ над текущими элементами $a_{2,2}$, $a_{2,k}$ и $a_{k,k}$.

(4в) Для каждого значения l от $k+1$ до n выполняем часть строчного преобразования, включающую $\cos \theta_{2,k}$ и $\sin \theta_{2,k}$ над $a_{2,l}$ и $a_{k,l}$.

(4г) Для каждого значения l от $k+1$ до n выполняем часть столбцового преобразования, включающую $\cos \theta_{2,l}$ и $\sin \theta_{2,l}$ над $a_{2,k}$ и $a_{k,l}$, используя тот факт, что $a_{2,k} = a_{k,2}$ в силу симметрии. После того как (4а), (4б), (4в) и (4г) для данного k окончены, будут выполнены преобразования первого основного шага над всеми элементами строки k и над элементами 3, 4, ..., k строки 2. Элементы 2, $k+1$, $k+2$, ..., n строки 2 подвергаются пока лишь преобразованиям, включающим $\theta_{2,3}$, $\theta_{2,4}$, ..., $\theta_{2,k}$.

(5) Переносим полученную k -ю строку во внешнюю память и возвращаемся к (4а) со следующим значением k .

25. Когда этапы (4) и (5) выполнены для всех соответствующих значений k , то тем самым выполнены и все вычисления над строкой 2. Измененные элементы первой строки (т. е. первые два элемента трехдиагональной матрицы), $\cos \theta_{2,k}$ и $\sin \theta_{2,k}$ (k от 3 до n) могут быть записаны во внешнюю память, а строка 2 перенесена в первую группу рабочих ячеек. Так как вторая строка на втором основном шаге играет такую же роль, как первая на первом шаге, то уже все готово для этапа (2) второго основного шага. Этап (1) на всех последующих основных шагах не нужен, так как соответствующая строка уже находится в быстродействующей памяти.

Следует подчеркнуть, что процесс, который мы только что описали, есть просто перестроенный процесс Гивенса. Хотя, прежде чем закончатся более ранние преобразования, выполняются части более поздних преобразований, элементы, затронутые ими, не влияют на завершение первых преобразований. Что касается числа арифметических операций и ошибок округления, то эти две схемы эквивалентны, но число переносов из внешней памяти и обратно во второй схеме существенно меньше.

Удобно для матрицы отвести во внешней памяти n^2 ячеек, сохраняя каждую строку полностью, несмотря на то, что поддиагональные элементы не используются, так как полное число ячеек необходимо для запоминания синусов и косинусов. Последние должны быть сохранены, если требуются собственные векторы. Удобный способ запоминания окончательной информации показан в (25.1). С этим способом запоминания «адресная»

арифметика особенно проста:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cos \theta_{2,3} & \cos \theta_{2,4} & \cos \theta_{2,5} \\ x & a_{2,2} & a_{2,3} & \cos \theta_{3,4} & \cos \theta_{3,5} \\ \sin \theta_{4,5} & x & a_{3,3} & a_{3,4} & \cos \theta_{4,5} \\ \sin \theta_{3,4} & \sin \theta_{3,5} & x & a_{4,4} & a_{4,5} \\ \sin \theta_{2,3} & \sin \theta_{2,4} & \sin \theta_{2,5} & x & a_{5,5} \end{bmatrix}. \quad (25.1)$$

Йогансен отметил, что если переносы из внешней памяти выполняются по k слов за один раз, то в быстродействующей памяти можно иметь $3n + k$ рабочих ячеек вместо $4n$, так как нам не нужно иметь в памяти одновременно целую строку матрицы A .

Анализ ошибок процесса Гивенса в арифметике с плавающей запятой

26. Чтобы закончить решение исходной проблемы собственных значений, нужно решить такую же проблему для трехдиагональной матрицы. Так как наш следующий метод также приводит к симметричной трехдиагональной матрице, оставим на некоторое время этот вопрос и дадим анализ ошибок приведения к трехдиагональному виду.

Очевидно, что преобразование Гивенса полностью охватывается общим анализом, данным в гл. 3, §§ 20—36, даже при таком вычислении углов, как описано здесь. Если мы обозначим *вычисленную* трехдиагональную матрицу через C , а *точное* произведение $1/2(n-1)(n-2)$ *точных* вращений, соответствующих *вычисленным* A_k , через R , то из главы 3 неравенства (27.2) имеем

$$\|C - RA_0 R^T\|_E \leq 12 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{4n-7} \|A_0\|_E \quad (26.1)$$

для стандартных вычислений с плавающей запятой.

Заметим, что вращения, входящие в процесс Гивенса, не составляют *полную совокупность* вращений, описанных в гл. 3, § 22, так как не требуется ни одно вращение, затрагивающее первую плоскость. Ясно, что оценка, приведенная там, тем более пригодна для меньшей совокупности вращений.

Оценка (26.1) не использует тот факт, что преобразованные матрицы имеют возрастающее число нулевых элементов, и в действительности учесть это совсем не просто. Однако самое большее, что можно ожидать от такого анализа, так это некоторое уменьшение множителя 12 в (26.1). Во всяком случае мы имеем

$$\left[\frac{\sum (\mu_i - \lambda_i)^2}{\sum \lambda_i^2} \right]^{1/2} \leq 12 \cdot 2^{-t} n^{3/2} (1 + 6 \cdot 2^{-t})^{4n-7}, \quad (26.2)$$

где μ_i суть собственные значения вычисленной трехдиагональной матрицы. Эту оценку можно сравнить с оценкой (18.3) для метода Якоби. Покажем в §§ 40 и 41, что ошибки, сделанные при вычислении собственных значений трехдиагональной матрицы, незначительны по сравнению с оценкой (26.2).

Анализ ошибок в арифметике с фиксированной запятой

27. Общий анализ в арифметике с фиксированной запятой из гл. 3, §§ 29—35 также непосредственно применяется к преобразованию Гивенса, но теперь очень легко учесть нули, которые постепенно получаются в процессе преобразования. В течение r -го основного шага имеем дело с матрицей порядка $(n - r + 1)$, а не n . Следовательно, в оценке (35.17) из главы 3 мы должны заменить $(2n - 4)^{1/2}$ на $(2n - 2r - 2)^{1/2}$. С учетом этого изменения имеем

$$\begin{aligned} \|C - RA_0R^T\|_2 &\leq 2^{-t-1} \sum_r [(2n - 2r - 2)^{1/2} + 9,0](n - r - 1) < \\ &< 2^{-t-1} \left(\frac{2^{3/2}}{5} n^{5/2} + 4,5n^2 \right) = x, \end{aligned} \quad (27.1)$$

при условии, что A_0 нормирована так, что

$$\|A_0\|_2 \leq 1 - x. \quad (27.2)$$

Аналогично можем показать, что

$$\|C - RA_0R^T\|_E \leq 2^{-t-1} \left(\frac{4}{5} n^{5/2} + 6n^2 \right) = y \quad (27.3)$$

при условии, что A_0 нормирована так, что

$$\|A_0\|_E < 1 - y. \quad (27.4)$$

Множители, связанные с членами n^2 в (27.1) и (27.3), велики и, возможно, при более тщательном анализе могут быть существенно уменьшены, но для достаточно больших n оценка в любом случае определяется членом $n^{5/2}$.

Численный пример

28. Здесь мы даем простой пример реализации метода Гивенса, который специально построен для того, чтобы проиллюстрировать природу численной устойчивости процесса. Мы уже не раз старались подчеркнуть, что окончательная трехдиагональная матрица очень близка к матрице, которая могла бы быть получена применением к A_0 *точных* ортогональных преобразований, соответствующих последовательно *вычисленным* A_r . Мы не утверждаем, что трехдиагональная матрица близка к той матрице, которая могла бы быть получена при повсеместном точном вычислении, но этот факт несуществен, если только собственные значения и собственные векторы, полученные нами, близки к собственным значениям и собственным векторам матрицы A_0 .

В табл. 4 сначала даем приведение к трехдиагональному виду матрицы пятого порядка, используя вычисления с фиксированной запятой и работа с пятью десятичными знаками. После получения нуля в позиции (1, 5) элементы в позициях (2, 3), (2, 4), (2, 5) в результате сокращения становятся очень малыми. Таким образом, эти элементы имеют большую относительную ошибку по сравнению с теми элементами, которые могли бы быть получены при точном вычислении. Следовательно, когда мы определяем вращение в плоскостях (3, 4) и (3, 5), углы будут сильно отличаться от значений, соответствующих точным вычислениям.

Таблица 4

$$A_0 = \begin{bmatrix} 0,71235 & 0,33973 & 0,28615 & 0,30388 & 0,29401 \\ & 1,18585 & -0,21846 & -0,06685 & -0,37360 \\ & & 0,18159 & 0,27955 & 0,38898 \\ & & & 0,23195 & 0,20496 \\ & & & & 0,46004 \end{bmatrix}$$

*Преобразование, использующее 5 десятичных знаков*Матрица A_3 , полученная после введения нулей в первую строку

$$A_3 = \begin{bmatrix} 0,71235 & 0,61325 & 0 & 0 & 0 \\ & 0,61874 & 0,00002 & 0,00001 & 0,00001 \\ & & 0,81366 & 0,51237 & 0,61325 \\ & & & 0,21436 & 0,21537 \\ & & & & 0,41267 \end{bmatrix}$$

Матрица A_5 , полученная после введения нулей во вторую строку

$$A_5 = \begin{bmatrix} 0,71235 & 0,61325 & 0 & 0 & 0 \\ & 0,61874 & 0,00002 & 0 & 0 \\ & & 1,48135 & 0,02405 & 0,11049 \\ & & & -0,07568 & -0,10328 \\ & & & & 0,03502 \end{bmatrix}$$

Окончательная трехдиагональная матрица A_6

$$A_6 = \begin{bmatrix} 0,71235 & 0,61325 & 0 & 0 & 0 \\ & 0,61874 & 0,00002 & 0 & 0 \\ & & 1,48135 & 0,11308 & 0 \\ & & & -0,01292 & 0,11694 \\ & & & & -0,02774 \end{bmatrix}$$

*Преобразование, использующее 6 десятичных знаков*Матрица A_3 , полученная после введения нулей в первую строку

$$A_3 = \begin{bmatrix} 0,712350 & 0,613256 & 0 & 0 & 0 \\ & 0,618749 & 0,000012 & 0,000007 & 0,000008 \\ & & 0,813654 & 0,512374 & 0,613255 \\ & & & 0,214360 & 0,215375 \\ & & & & 0,412665 \end{bmatrix}$$

Матрица A_5 , полученная после введения нулей во вторую строку

$$A_5 = \begin{bmatrix} 0,712350 & 0,613256 & 0 & 0 & 0 \\ & 0,618749 & 0,000016 & 0 & 0 \\ & & 1,486335 & -0,068499 & 0,024712 \\ & & & -0,079492 & -0,102483 \\ & & & & 0,033837 \end{bmatrix}$$

Окончательная трехдиагональная матрица A_6

$$A_6 = \begin{bmatrix} 0,712350 & 0,613256 & 0 & 0 & 0 \\ & 0,618749 & 0,000016 & 0 & 0 \\ & & 1,486335 & 0,072820 & 0 \\ & & & -0,001012 & -0,115055 \\ & & & & -0,044643 \end{bmatrix}$$

Собственные значения с 8 десятичными знаками

A_0	A_6 (5-значное вычисление)	A_6 (6-значное вычисление)
1,48991 259	1,48991 000	1,48991 239
1,28058 937	1,28057 855	1,28058 869
-0,14126 880	-0,14126 174	-0,14126 895
0,09203 628	0,09204 175	0,09203 657
0,05051 055	0,05051 145	0,05051 030

Мы подчеркнули это повторением вычислений с шестью десятичными знаками. Теперь элементы (2, 3), (2, 4) и (2, 5) соответственно равны

$$0,000012, \quad 0,000007, \quad 0,000008$$

по сравнению с прежними значениями

$$0,00002, \quad 0,00001, \quad 0,00001$$

для 5-значного вычисления. Например, для вращения в плоскости (3, 4) имеем $\operatorname{tg} \theta = 7/12$ и $\operatorname{tg} \theta = 1/2$ соответственно для обоих случаев. В результате обе трехдиагональные матрицы почти полностью отличаются друг от друга в последних строках и столбцах. Однако наш анализ ошибок показывает, что обе трехдиагональные матрицы должны иметь собственные значения, близкие к собственным значениям матрицы A_0 .

В табл. 4 мы дали собственные значения всех этих матриц с 8 десятичными знаками и они подтверждают наше утверждение. Заметим, что в обоих случаях влияние ошибок округления, сделанных во всем приведении, на каждое собственное значение несущественно (меньше одной единицы последнего знака). На практике мы установили, что действительные ошибки всегда намного меньше, чем те верхние оценки, которые мы получили, и грубые оценки, учитывающие статистическое распределение.

Заметим, что мы действительно имеем значительную свободу в выборе углов вращения в плоскостях (3, 4) и (3, 5). Например, при осуществлении вращения в плоскости (3, 4) мы должны выбрать $\cos \theta$ и $\sin \theta$ лишь так, чтобы они были точной парой с рабочей точностью и чтобы, например, в случае счета с шестью знаками выполнялось неравенство

$$|-(0,000012) \sin \theta + (0,000007) \cos \theta| < \frac{1}{2} \cdot 10^{-6}.$$

Методы, которые мы описали, таковы, что θ очень точно соответствует формуле $\operatorname{tg} \theta = \frac{0,000007}{0,000012}$, а это самое простое, что можно сделать. Однако для вычисления $\operatorname{tg} \theta$ одинаково хорошо подходило бы любое значение x/y , где

$$0,0000065 \leq x \leq 0,0000075, \quad 0,0000115 \leq y \leq 0,0000125,$$

хотя такие значения соответствуют большому разбросу в элементах трехдиагональных матриц.

Если это показалось неожиданным, давайте вернемся к точным вычислениям и предположим, что произошло полное сокращение и во время исключения элемента (1, 5) элементы (2, 3), (2, 4), (2, 5) стали нулями. Теперь *любые* точные вращения в плоскостях (3, 4) и (3, 5) не изменяли бы эти три элемента и могли бы рассматриваться как приближенные преобразования Гивенса. Различные вращения дали бы различные трехдиагональные матрицы, но все они были бы подобны A_0 .

Если в точном процессе происходит полное сокращение, то мы можем выбрать произвольными все знаки в соответствующих углах вращения.

Метод Хаусхолдера

29. В течение нескольких лет казалось, что процесс Гивенса, вероятно, должен быть самым эффективным методом приведения матрицы к трехдиагональному виду, но в 1958 г. Хаусхолдер высказал мысль, что это приведение можно выполнить более эффективно, используя матрицы

отражения вместо плоских вращений. В гл. 6, § 7 покажем, что преобразования Гивенса и Хаусхолдера в основном совпадают.

Это приведение состоит из $(n - 2)$ шагов, на r -м из которых получают нули в r -й строке и r -м столбце, причем сохраняются нули, полученные на предыдущих шагах. Структура элементов *перед* r -м шагом иллюстрируется схемой (29.1) для случая $n = 7, r = 4$

$$A_{r-1} = \begin{matrix} & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{matrix} \begin{matrix} r \\ \left\{ \begin{array}{cccc|ccc} \times & \times & & & & & \\ & \times & \times & \times & & & \\ & & \times & \times & \times & & \\ & & & \times & \times & \times & \times \\ \hline & & & & \times & \times & \times \\ & & & & \times & \times & \times \\ & & & & \times & \times & \times \end{array} \right\} \\ n-r \end{matrix} = \begin{bmatrix} C_{r-1} & 0 \\ 0 & b_{r-1} & B_{r-1} \end{bmatrix}. \quad (29.1)$$

Здесь b_{r-1} — вектор, имеющий $(n - r)$ компонент, C_{r-1} — трехдиагональная матрица порядка r и B_{r-1} — квадратная матрица порядка $(n - r)$. Матрица P_r r -го преобразования может быть представлена в виде

$$P_r = \begin{matrix} r \\ \left\{ \begin{array}{c|c} I & 0 \\ \hline 0 & Q_r \end{array} \right\} \\ n-r \end{matrix} = \begin{bmatrix} I & 0 \\ 0 & I - 2v_r v_r^T \end{bmatrix}, \quad (29.2)$$

где v_r — вектор единичной длины, имеющий $(n - r)$ компонент. Следовательно, имеем

$$A_r = P_r A_{r-1} P_r = \begin{matrix} r \\ \left\{ \begin{array}{c|c} C_{r-1} & 0 \\ \hline 0 & c_r & Q_r B_{r-1} Q_r \end{array} \right\} \\ n-r \end{matrix}, \quad (29.3)$$

где

$$c_r = Q_r b_{r-1}. \quad (29.4)$$

Если выбрать v_r так, что c_r будет нулевым, за исключением первой компоненты, то A_r в первых $(r + 1)$ строках и столбцах будет трехдиагональной. Итак, мы представили условие, определяющее P_r , в стандартной форме, рассмотренной в гл. 3, §§ 37, 38, но приведенные там результаты удобно переформулировать на языке новых обозначений. Обозначим текущий элемент (i, j) матрицы A_{r-1} через a_{ij} , опуская индекс r . Будем предполагать, что запоминается лишь верхний треугольник элементов, и представим все формулы в зависимости от их значений. Формулы становятся наиболее простыми, если мы представим матрицу P_r в виде $P_r = I - 2w_r w_r^T$, где w_r есть вектор единичной длины, имеющий n компонент, из которых первые r равны нулю. Тогда имеем

$$P_r = I - 2w_r w_r^T = I - u_r u_r^T / 2K_r^2, \quad (29.5)$$

где

$$\left. \begin{aligned} u_{ir} &= 0 \\ u_{r+1,r} &= a_{r,r+1} \mp S_r, & u_{ir} &= a_{ri} & (i = r+2, \dots, n), \\ S_r &= \left(\sum_{i=r+1}^n a_{ri}^2 \right)^{1/2}, & 2K_r^2 &= S_r^2 \mp a_{r,r+1} S_r. \end{aligned} \right\} \quad (29.6)$$

В уравнении, определяющем $u_{r+1, r}$, знак S_r берется равным знаку $a_{r, r+1}$, так что при сложении не может иметь место сокращение. Новое значение $a_{r, r+1}$ равно $\pm S_r$. (Мы сравним это с методом Гивенса, для которого все внедиагональные элементы трехдиагональной матрицы, полученные по нашим формулам, неотрицательны.) Заметим, что если все элементы, которые должны быть исключены на данном шаге, уже равны нулю, то матрица преобразования не будет единичной, как можно было бы ожидать. (Действительно, I не является матрицей отражения согласно нашему определению.) Матрица преобразования отличается от единичной тем, что $(r+1)$ -й диагональный элемент равен -1 вместо $+1$. Мы можем договориться, что в этом случае будем опускать соответствующее преобразование.

Использование симметрии

30. Важно полностью использовать симметрию, но для этого при вычислении A_r по формуле

$$A_r = P_r A_{r-1} P_r, \quad (30.1)$$

требуется более точная формулировка. Мы имеем в действительности

$$A_r = (I - u_r u_r^T / 2K_r^2) A_{r-1} (I - u_r u_r^T / 2K_r^2), \quad (30.2)$$

и если обозначим

$$A_{r-1} u_r / 2K_r^2 = p_r, \quad (30.3)$$

то уравнение (30.2) может быть записано в виде

$$A_r = A_{r-1} - u_r p_r^T - p_r u_r^T + u_r (u_r^T p_r) u_r^T / 2K_r^2. \quad (30.4)$$

Следовательно, если введем q_r , определенное равенством

$$q_r = p_r - \frac{1}{2} u_r (u_r^T p_r / 2K_r^2), \quad (30.5)$$

то имеем

$$A_r = A_{r-1} - u_r q_r^T - q_r u_r^T, \quad (30.6)$$

и теперь правая часть удобна для использования симметрии. Нас интересует лишь матрица порядка $(n-r)$ в нижнем правом углу A_r , так как элементы в первых $(r-1)$ строках и столбцах те же, что у матрицы A_{r-1} , и преобразование P_r построено для того, чтобы исключить a_{ri} ($i = r+1, \dots, n$) и заменить $a_{r, r+1}$ на $\pm S_r$.

Первые $(r-1)$ компонент вектора p_r равны нулю, и нам не нужен r -й элемент. Следовательно, для вычисления этого вектора требуется $(n-r)^2$ умножений. Число операций, требующееся для вычисления q_r , порядка $(n-r)$ и относительно мало. Наконец, для вычисления соответствующих элементов A_r по формуле (30.6) требуется около $(n-r)^2$ умножений, так как нам нужно вычислить лишь верхний треугольник элементов. Поэтому общее число умножений во всем преобразовании приблизительно равно $\frac{2}{3} n^3$, по сравнению с $\frac{4}{3} n^3$ в преобразовании Гивенса.

В процессе, который мы дали, около n извлечений квадратных корней, в процессе Гивенса их $\frac{1}{2} n^2$.

Некоторые соображения о необходимой величине памяти

31. Если нам нужны собственные векторы, то должны быть сохранены элементы матриц преобразования. Очевидно, что достаточно хранить ненулевые элементы векторов u_r и информацию, необходимую для получения $2K_r^2$. Вектор u_r имеет $(n - r)$ ненулевых элементов, тогда как он исключает только $(n - r - 1)$ элементов матрицы A_{r-1} . Элемент $(r, r + 1)$ не становится нулем и в действительности является внедиагональным элементом трехдиагональной матрицы. Но эти внедиагональные элементы удобно хранить отдельно, и, следовательно, мы можем записать все ненулевые элементы u_r на месте исходной матрицы. Это особенно удобно, так как $u_{ir} = a_{ri}$ ($i = r + 2, \dots, n$), и поэтому все элементы u_r , за исключением $u_{r+1, r}$, уже находятся в нужных позициях.

Все последовательные p_r могут запоминаться в одной и той же группе из n рабочих ячеек, и мы можем каждое q_r записывать на место p_r . Так как в q_r и p_r есть лишь $(n - r)$ интересующих нас элементов, то и наддиагональные элементы $\pm S_r$ мы можем запоминать по мере их получения в той же группе из n рабочих ячеек. Если требуется, то из $\pm S_r$ и $u_{r+1, r}$ можно получить $2K_r^2$, используя соотношение

$$2K_r^2 = S_r^2 \mp a_{r, r+1} S_r = S_r (S_r \mp a_{r, r+1}) = -(\pm S_r) u_{r+1, r}. \quad (31.1)$$

Итак, при сохранении элементов матриц преобразования данный процесс требует только $\frac{1}{2}(n^2 + 3n)$ рабочих ячеек по сравнению с n^2 , необходимых в методе Гивенса.

Процесс Хаусхолдера на вычислительной машине с двухступенчатой памятью

32. Если метод Хаусхолдера осуществляется таким способом на машине с двухступенчатой памятью, то на r -м основном шаге необходимо сделать два переноса наддиагональных элементов матрицы B_{r-1} (ср. с (29.1)) в быстродействующую память, один раз для вычисления p_r и второй раз для получения элементов A_r . В модифицированной схеме метода Гивенса, описанной в § 24, на каждом основном шаге требуется лишь один такой перенос, но та же экономия может быть достигнута здесь следующим образом.

Во время вычисления A_r , которое включает в себя считывание A_{r-1} из внешней памяти и замену ее на A_r , также могут быть вычислены u_{r+1} и p_{r+1} . A_r отличается от A_{r-1} лишь в строках и столбцах от r до n , и как только становится известной $(r + 1)$ -я строка матрицы A_r , может быть вычислен вектор u_{r+1} . По мере того как определяется каждый элемент A_r , он может быть использован для вычисления соответствующего вклада в вектор $A_r u_{r+1}$.

Так как мы предполагаем, что работаем только с наддиагональными элементами, то вычисленный элемент (i, j) матрицы A_r должен быть использован для определения соответствующего вклада как в $(A_r u_{r+1})_i$, так и в $(A_r u_{r+1})_j$. Когда вычисление A_r и ее запись во внешнюю память закончены, то будут определены u_{r+1} и $A_r u_{r+1}$, а, следовательно, $2K_{r+1}^2$, p_{r+1} и q_{r+1} могут быть вычислены без дальнейшего обращения к внешней памяти. Этот способ делает метод Хаусхолдера столь же эффективным, как и метод Гивенса.

Если нам нужно полностью реализовать точность, достигаемую накоплением скалярных произведений при вычислении вектора $A_r u_{r+1}$, то каждый его элемент должен удерживаться с двойной точностью и получаться постепенно по мере вычисления A_r . Итак, в рабочей памяти должны быть размещены пять векторов u_r , q_r , u_{r+1} , $A_r u_{r+1}$ и текущая строка A_{r-1} , которая преобразуется в строку A_r . Так как $A_r u_{r+1}$ временно находится с двойной точностью, то требуется $6n$ рабочих ячеек. Если из внешней памяти и обратно переносится по k слов за один раз, то число рабочих ячеек может быть снижено до $5n + k$, так как целая текущая строка матрицы A_{r-1} нам не нужна в памяти одновременно.

Метод Хаусхолдера в арифметике с фиксированной запятой

33. Метод, описанный нами в §§ 29, 30, не пригоден в том виде для работы с фиксированной запятой. При работе с плавающей запятой член $(u_r u_r^T)/2K_r^2$ эффективно представляется в виде $u_r (u_r^T/2K_r^2)$. Если элементы, которые определяют текущее преобразование, очень малы, то компоненты вектора $u_r/2K_r^2$ будут много больше единицы. Уилкинсон (1960) дал другую формулу, где эффективно используется представление $2(u_r/2K_r)(u_r^T/2K_r)$, хотя при этом на каждом преобразовании нужно извлекать еще один квадратный корень. Для матриц высокого порядка время, затрачиваемое на извлечение дополнительных квадратных корней, незначительно по сравнению с общим временем счета, и кроме этого, в данной формуле легко удовлетворяются требования нормировки. Однако существуют формулы, в которых нет дополнительных извлечений квадратных корней, и сейчас мы опишем одну из них отчасти потому, что она нам потребуется в следующей главе, и отчасти потому, что ее анализ ошибок имеет некоторые поучительные особенности.

Вместо того чтобы работать непосредственно с u_r , воспользуемся нормированным вектором

$$y_r = u_r / u_{r+1, r}. \quad (33.1)$$

Из (29.5) и (29.6) имеем

$$\left. \begin{aligned} y_{ir} &= 0 & (i = 1, 2, \dots, r), \\ y_{r+1, r} &= 1, \quad y_{ir} = a_{ri} / (a_{r, r+1} \mp S_r) & (i = r+2, \dots, n), \end{aligned} \right\} \quad (33.2)$$

и

$$P_r = I - u_r u_r^T / 2K_r^2 = I - \left(\frac{S_r \mp a_{r, r+1}}{S_r} \right) y_r y_r^T. \quad (33.3)$$

Теперь, если все a_{ri} ($i = r+1, \dots, n$) малы, то вычисленное S_r обычно будет иметь сравнительно большую относительную ошибку. Тем не менее очевидно, что все вычисленные элементы вектора y_r будут ограничены по модулю единицей, хотя вычисленный y_r может существенно отличаться от вектора, точно соответствующего заданным a_{ri} . Для того чтобы быть уверенными в ортогональности матрицы преобразования, мы должны взять

$$P_r = I - 2y_r y_r^T / \|y_r\|^2 \quad (33.4)$$

место правой части (33.3).

Теперь A_r можно вычислить в такой последовательности:

$$p_r = A_{r-1} y_r / \|y_r\|^2, \quad (33.5)$$

$$q_r = p_r - (y_r^T p_r) y_r / \|y_r\|^2, \quad (33.6)$$

$$A_r = A_{r-1} - 2y_r q_r^T - 2q_r y_r^T. \quad (33.7)$$

Так как $1 \leq \|y_r\|^2 \leq 2$, то с различных точек зрения необходимо ввести в эту формулу множитель $1/2$. Более удобный путь действия в некоторой степени зависит от особенностей фиксированной запятой, однако этот вопрос мы разбирать не будем.

Здесь мы имеем другую ситуацию по сравнению с той, к которой привыкли в нашем анализе ошибок. *Вычисленная матрица преобразования может значительно отличаться от той, которая соответствует точному вычислению по уже полученной матрице A_{r-1} .* Но это совсем неважно. В силу того, что $y_{r+1, r} = 1$, вычисленная величина $\|y_r\|^2$ имеет малую относительную ошибку и, следовательно, матрица, полученная из (33.4), почти ортогональна; далее, элементы матрицы $P_r A_{r-1} P_r$ в соответствующих позициях с рабочей точностью равны нулю. Это единственные требования, которые мы должны выполнить.

Численный пример

34. Эти соображения иллюстрируем примером из табл. 5, который был вычислен с использованием формул § 33. Все элементы матрицы A_0 , определяющие первое преобразование, очень малы. Поскольку нормировка не используется, вычисленное значение S_1 имеет большую относительную ошибку. В действительности вычисленное значение есть 0,00005, тогда как верное значение равно 0,000053852... Если бы мы должны были связать

Таблица 5

A_0					
$\begin{bmatrix} 0,23157 & 0,00002 & 0,00003 & 0,00004 \\ 0,00002 & 0,31457 & 0,21321 & 0,31256 \\ 0,00003 & 0,21321 & 0,18175 & 0,21532 \\ 0,00004 & 0,31256 & 0,21532 & 0,41653 \end{bmatrix}$				$S_1 = 0,00005$	
				$\ y_1\ ^2 = 1,51020$	
				$y_1^T p_1 / \ y_1\ ^2 = 0,49519$	
				$y_1^T = (0,00000, \quad 1,00000, \quad 0,42857, \quad 0,57143)$	
				$p_1^T = (0,00004, \quad 0,38707, \quad 0,27423, \quad 0,42568)$	
				$q_1^T = (0,00004, \quad -0,10812, \quad 0,06201, \quad 0,14271)$	
A_1					
$\begin{bmatrix} 0,23157 & -0,00005 & 0,00000 & 0,00000 \\ -0,00005 & 0,74705 & 0,18186 & 0,15071 \\ 0,00000 & 0,18186 & 0,07545 & 0,02213 \\ 0,00000 & 0,15071 & 0,02213 & 0,09033 \end{bmatrix}$				$S_2 = 0,23619$	
				$\ y_2\ ^2 = 1,12997$	
				$y_2^T p_2 / \ y_2\ ^2 = 0,08078$	
				$y_2^T = (0,00000, \quad 0,00000, \quad 1,00000, \quad 0,36051)$	
				$p_2^T = (0,00000, \quad 0,20903, \quad 0,07383, \quad 0,04840)$	
				$q_2^T = (0,00000, \quad 0,20903, \quad -0,00695, \quad 0,01928)$	
A_2					
$\begin{bmatrix} 0,23157 & -0,00005 & 0,00000 & 0,00000 \\ -0,00005 & 0,74705 & -0,23619 & 0,00000 \\ 0,00000 & -0,23619 & 0,10325 & -0,01142 \\ 0,00000 & 0,00000 & -0,01142 & 0,06253 \end{bmatrix}$					

Продолжение табл. 5

Точное вычисление первого шага

$$S_1 = 0,000053852, \quad \|y_1\|^2 = 1,45837, \quad y_1^T p_1 / \|y_1\|^2 = 0,50464$$

$$\begin{aligned} y_1^T &= (0,00000, \quad 1,00000, \quad 0,40622, \quad 0,54162) \\ p_1^T &= (0,00004, \quad 0,39117, \quad 0,27679, \quad 0,42899) \\ q_1^T &= (0,00004, \quad -0,11347, \quad 0,07180, \quad 0,15537) \end{aligned}$$

$$A_1$$

$$\begin{bmatrix} 0,23157 & -0,00005 & 0,00000 & 0,00000 \\ & 0,76845 & 0,16180 & 0,12414 \\ & & 0,06508 & 0,01107 \\ & & & 0,07927 \end{bmatrix}$$

с $y_1 y_1^T$ множитель $(S_1 + a_{1,2})/S_1$ так, как указано в (33.3), то матрица преобразования даже приближенно не была бы ортогональной. Использование формулы (33.4) гарантирует почти точную ортогональность и полностью оправдано то, что мы взяли соответствующие элементы матрицы A_1 равными нулю. Однако A_1 сильно отличается от той матрицы, которая могла бы быть получена с использованием плавающей запятой и формул § 30. Эти формулы дают ортогональную матрицу, которая близка к точной матрице, соответствующей A_0 .

В конце таблицы мы даем вектор y_1 , соответствующий точному вычислению, и соответствующую матрицу A_1 . Видно, что эти y_1 и A_1 значительно отличаются от вычисленных.

Анализ ошибок метода Хаусхолдера

35. Был дан целый ряд различных анализов ошибок метода Хаусхолдера. Наш общий анализ ошибок из гл. 3, § 45 нуждается лишь в небольшом изменении для того, чтобы охватить симметричный случай. Уравнение (45.5) гл. 3 показывает, что если C есть вычисленная трехдиагональная матрица, полученная с использованием накопления с плавающей запятой, и R есть точное произведение ортогональных матриц, соответствующих вычисленным A_r , то

$$\|C - RA_0 R^T\|_E \leq 2K_1 n 2^{-t} (1 + K_1 2^{-t})^{2n} \|A_0\|_E \quad (35.1)$$

для некоторой константы K_1 . Для операций с плавающей запятой, использующихся на АСЕ, мы показали, что $2K_1$ меньше 40. Следовательно, в обозначениях § 26 имеем

$$\left[\frac{\sum (\mu_i - \lambda_i)^2}{\sum \lambda_i^2} \right]^{1/2} \leq 40 n 2^{-t} (1 + 20 \cdot 2^{-t})^{2n}, \quad (35.2)$$

и, как обычно, последний множитель в правой части (35.2) не имеет значения в приемлемых пределах этой оценки. Первый результат такого вида был получен Ортега (1963), причем его доказательство существенно отличается от нашего. Как мы отмечали в гл. 3, § 45, множитель $n 2^{-t}$ едва ли может быть улучшен, так как он получается даже тогда, когда мы предполагаем, что каждое преобразование выполняется *точно*, и только перед переходом к следующему шагу полученная матрица A_r округляется до t двоичных разрядов.

Ортега получил также оценку, соответствующую обычным вычислениям с плавающей запятой. Его оценка, согласно нашим обозначениям, по существу имеет такой вид:

$$\left[\frac{\sum (\mu_i - \lambda_i)^2}{\sum \lambda_i^2} \right]^{1/2} \leq 2K_2 n^2 2^{-t} (1 + K_2 n 2^{-t})^{2n}. \quad (35.3)$$

Константы K_1 и K_2 зависят от особенностей рассматриваемой арифметики, и тщательный их учет в анализе может несколько уменьшить полученные значения. Но, сравнивая (26.2) и (35.3), мы видим, что в пределах использования имеем оценки порядка $n^{3/2} 2^{-t}$ и $n^2 2^{-t}$ соответственно для методов Гивенса и Хаусхолдера, выполненных в обычной арифметике с плавающей запятой. Поэтому при больших n оценка для процесса Гивенса лучше. Несомненно оценка (35.3) слабее, чем большинство оценок, которые мы получили, но это еще не показывает, какой процесс дает более точные результаты на практике. Трудность заключается в том, что вследствие статистического распределения ошибок округления собственные значения, полученные на практике, очень точны. Поэтому для получения важной информации об относительной точности обоих методов нужно исследовать очень большое число матриц высокого порядка.

Хотя верхняя оценка (35.2) для метода Хаусхолдера получена при использовании вычислений с плавающей запятой и накоплением, все же можно предположить, что она несколько завышена, и для матриц, которые мы подробно исследовали, обычно не достигалось даже значение $n^{1/2} 2^{-t}$. Если мы будем считать значение $n^{1/2} 2^{-t}$ как верное, то очевидно, что влияние ошибок округления, вероятно, не должно быть ограничивающим фактором на ближайшее будущее. Если мы возьмем, например, $t = 40$ и $n = 10^4$, то $n^{1/2} 2^{-t}$ приблизительно равно 10^{-10} . Следовательно, мы могли бы ожидать, что

$$\left[\frac{\sum (\lambda_i - \mu_i)^2}{\sum \lambda_i^2} \right]^{1/2} < 10^{-10}, \quad (35.4)$$

а это очень хороший результат. Но полная матрица порядка 10^4 имеет 10^8 элементов, и ее преобразование требует около $2/3 10^{12}$ умножений. Даже со скоростью выполнения умножений в 1 мксек время счета составило бы более 200 часов.

Детальный анализ ошибок метода Хаусхолдера в арифметике с фиксированной запятой и накоплением был дан Уилкинсоном (1962b). Этот анализ показал, что

$$\max |\lambda_i - \mu_i| < K_3 n^2 2^{-t} \quad (35.5)$$

при таком условии нормировки A_0 :

$$\|A_0\|_2 < \frac{1}{2} - K_3 n^2 2^{-t}. \quad (35.6)$$

Эта оценка была получена для формул, имеющих два квадратных корня на каждом основном шаге, но при условии, что накапливаются скалярные произведения; несколько других формул дают результаты типа (35.5),

но с различными значениями K_3 . Снова опыт позволяет высказать предположение, что в общем случае даже $n2^{-l}$ является существенной переоценкой.

Для вычислений с фиксированной запятой без накопления оценка вида

$$\max |\lambda_i - \mu_i| < K_4 n^{5/2} 2^{-l} \quad (35.7)$$

является самой лучшей среди тех, которые до сих пор были получены.

Собственные значения симметричной трехдиагональной матрицы

36. Для того чтобы завершить метод Гивенса или Хаусхолдера, мы должны решить проблему собственных значений для симметричной трехдиагональной матрицы. Трехдиагональные матрицы обычно возникают как исходные данные во многих физических задачах и, следовательно, имеют самостоятельное значение. Так как в каждой такой матрице только $(2n - 1)$ независимых элементов, то можно ожидать, что эта проблема будет значительно проще, чем для исходной матрицы.

Рассмотрим теперь симметричную трехдиагональную матрицу C и для удобства обозначим

$$c_{ii} = \alpha_i, \quad c_{i, i+1} = c_{i+1, i} = \beta_{i+1}. \quad (36.1)$$

Если r элементов из β_i равны нулю, то C есть прямая сумма $(r + 1)$ трехдиагональных матриц меньшего порядка, т. е.

$$C = \begin{bmatrix} C^{(1)} & & & \\ & C^{(2)} & & \\ & & \ddots & \\ & & & C^{(r+1)} \end{bmatrix}, \quad (36.2)$$

где $C^{(s)}$ имеет порядок m_s и $\sum m_s = n$. В совокупности собственные значения матриц $C^{(s)}$ те же, что у матрицы C , и если x есть собственный вектор матрицы $C^{(s)}$, то соответствующий собственный вектор y матрицы C таков:

$$y^T = (\underbrace{0}_{m_1}, \underbrace{0}_{m_2}, \dots, \underbrace{0}_{m_{s-1}}, \underbrace{x^T}_{m_s}, \underbrace{0}_{m_{s+1}}, \dots, \underbrace{0}_{m_{r+1}}). \quad (36.3)$$

Следовательно, не теряя общности, можно предположить, что ни один из β_i не равен нулю.

Свойство последовательности Штурма

37. Если мы обозначим главный минор порядка r матрицы $(C - \lambda I)$ через $p_r(\lambda)$, то, определяя $p_0(\lambda)$ равным единице, имеем

$$p_1(\lambda) = \alpha_1 - \lambda, \quad (37.1)$$

$$p_i(\lambda) = (\alpha_i - \lambda)p_{i-1}(\lambda) - \beta_i^2 p_{i-2}(\lambda) \quad (i = 2, \dots, n). \quad (37.2)$$

Нули $p_r(\lambda)$ являются собственными значениями подматрицы главного минора порядка r матрицы C , и следовательно, они вещественные. Обозначим эту матрицу через C_r .

Теперь мы знаем, что собственные значения каждой C_r разделяют собственные значения C_{r+1} по крайней мере в слабом смысле (гл. 2, § 47).

Покажем, что если ни один из β_i не равен нулю, то мы должны иметь строгое разделение. Если μ есть нуль полиномов $p_r(\lambda)$ и $p_{r-1}(\lambda)$, то из (37.2) для $i = r$ и предположения, что β_r не равно нулю, получаем, что μ есть также нуль полинома $p_{r-2}(\lambda)$. Следовательно, из (37.2) для $i = r - 1$ вытекает, что μ является нулем $p_{r-3}(\lambda)$. Продолжая таким же путем, находим, что μ есть нуль $p_0(\lambda)$, а так как $p_0(\lambda) = 1$, то мы пришли к противоречию.

Отсюда заключаем, что если симметричная матрица имеет собственное значение кратности k , то трехдиагональная матрица, полученная точным выполнением метода Гивенса или Хаусхолдера, должна иметь по крайней мере $(k - 1)$ нулевых наддиагональных элементов. С другой стороны, присутствие нулевого наддиагонального элемента в трехдиагональной матрице не означает присутствия кратных корней.

38. Строгое разделение нулей симметричных трехдиагональных матриц с ненулевыми наддиагональными элементами составляет основу одного из самых эффективных методов определения собственных значений (Гивенс, 1954). Сначала докажем следующее предложение.

Пусть величины $p_0(\mu)$, $p_1(\mu)$, ..., $p_n(\mu)$ вычисляются для некоторого значения μ . Тогда $s(\mu)$, число совпадений знаков в последовательных членах этой последовательности, равно числу тех собственных значений матрицы C , которые строго больше, чем μ .

Для того чтобы высказанное выше утверждение имело смысл для всех значений μ , мы должны установить знак нулевого значения $p_r(\mu)$. Если $p_r(\mu) = 0$, то $p_r(\mu)$ должно быть взято с противоположным знаком по сравнению с $p_{r-1}(\mu)$. (Заметим, что никакие два последовательных $p_i(\mu)$ не могут быть равны нулю.)

Доказательство по индукции. Предположим, что число совпадений s в последовательности $p_0(\mu)$, $p_1(\mu)$, ..., $p_r(\mu)$ равно числу тех собственных значений матрицы C_r (т. е. числу нулей $p_r(\lambda)$), которые больше чем μ . Тогда мы можем обозначить собственные значения C_r через $x_1, x_2, \dots, x_r$, где

$$x_1 > x_2 > \dots > x_s > \mu \geq x_{s+1} > x_{s+2} > \dots > x_r, \quad (38.1)$$

и нет ни одного совпадающего собственного значения, так как β_i не равны нулю. Собственные значения $y_1, y_2, \dots, y_{r+1}$ матрицы C_{r+1} разделяют собственные значения матрицы C_r в строгом смысле, и следовательно,

$$y_1 > x_1 > y_2 > x_2 > \dots > y_s > x_s > y_{s+1} > x_{s+1} > \dots > y_r > x_r > y_{r+1}. \quad (38.2)$$

Это показывает, что или s , или $s + 1$ собственных значений C_{r+1} больше чем μ . Очевидно, мы имеем

$$p_r(\mu) = \prod_{i=1}^r (x_i - \mu) \quad \text{и} \quad p_{r+1}(\mu) = \prod_{i=1}^{r+1} (y_i - \mu). \quad (38.3)$$

Если ни одно из x_i и y_i не равно μ , то наш результат устанавливается немедленно. В этом случае, если $y_{s+1} > \mu$, то $p_r(\mu)$ и $p_{r+1}(\mu)$ имеют один и тот же знак, и, следовательно, в увеличенном ряде на одно совпадение больше; если же $y_{s+1} < \mu$, то $p_r(\mu)$ и $p_{r+1}(\mu)$ имеют различные знаки, поэтому увеличенная последовательность имеет то же самое число совпадений. В любом случае мы имеем соответствие с числом собственных значений, больших чем μ .

Если $y_{s+1} = \mu$, то мы должны взять знак $p_{r+1}(\mu)$ противоположным знаком $p_r(\mu)$ (которое не может быть в данном случае равно нулю), так что увеличенный ряд имеет число совпадений, соответствующее числу собственных значений, строго больших μ .

Если $x_{s+1} = \mu$, то должна быть та ситуация, которая иллюстрируется рис. 2 для случая $r = 5, s = 2$. Знак нулевого значения $p_r(\mu)$ должен быть взят противоположным знаком $p_{r-1}(\mu)$. Из строгого разделения нулей $p_{r-1}(\lambda)$, $p_r(\lambda)$ и $p_{r+1}(\lambda)$ следует, что по сравнению с $p_{r-1}(\lambda)$ полином $p_{r+1}(\lambda)$ должен иметь точно на один больше тех нулей, которые больше μ , и точно на один больше тех нулей, которые меньше μ . Следовательно, $p_{r+1}(\mu)$ должно иметь противоположный знак по сравнению со знаком $p_{r-1}(\mu)$, и поэтому тот же знак, который мы предположили у $p_r(\mu)$.

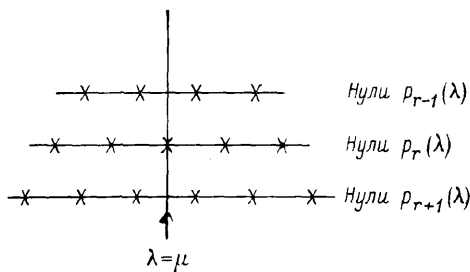


Рис. 2.

Таким образом, число совпадений увеличилось на одно, и это соответствует тому, что число нулей, больших μ , также увеличилось на 1.

Читателю следует проверить, что если нас интересуют только собственные значения самой матрицы C , то знаки, связанные с нулевыми значениями $p_r(\mu)$ ($r = 1, \dots, n-1$), не играют никакой роли. Нужно лишь придерживаться строгого правила в отношении $p_r(\mu)$.

Мы можем сказать, что полиномы $p_0(\lambda)$, $p_1(\lambda)$, $\dots$, $p_n(\lambda)$ обладают свойством последовательности Штурма.

Метод деления пополам

39. Свойство последовательности Штурма может быть использовано для локализации любого собственного значения, например, k -го в порядке уменьшения, независимо от остальных. Предположим, что у нас есть два значения a_0 и b_0 таких, что

$$b_0 > a_0, \quad s(a_0) \geq k, \quad s(b_0) < k, \quad (39.1)$$

тогда мы знаем, что λ_k лежит в интервале (a_0, b_0) . Теперь мы можем локализовать λ_k в интервале (a_p, b_p) длиной в $(b_0 - a_0)/2^p$ за p шагов итерационного процесса, r -й шаг которого состоит в следующем. Пусть

$$c_r = \frac{1}{2} (a_{r-1} + b_{r-1})$$

есть середина интервала (a_{r-1}, b_{r-1}) . Тогда вычисляем последовательность

$$p_0(c_r), p_1(c_r), \dots, p_n(c_r)$$

и затем определяем $s(c_r)$. Если $s(c_r) \geq k$, то берем $a_r = c_r$ и $b_r = b_{r-1}$; если $s(c_r) < k$, то берем $a_r = a_{r-1}$ и $b_r = c_r$. В любом случае

$$s(a_r) \geq k \quad \text{и} \quad s(b_r) < k, \quad (39.2)$$

так что λ_k находится в интервале (a_r, b_r) . Подходящими значениями для a_0 и b_0 являются $\pm \|C\|_\infty$, и мы имеем $\|C\|_\infty = \max_i \{ |\beta_i| + |\alpha_i| + |\beta_{i+1}| \}$.

Численная устойчивость метода деления пополам

40. При точном вычислении метод § 39 мог бы быть использован для определения любого собственного значения с любой наперед заданной точностью. На практике ошибки округления будут ограничивать достижимую точность, но тем не менее, как мы установим, метод исключительно устойчив. Мы дадим анализ ошибок, соответствующий использованию стандартной арифметики с плавающей запятой.

Покажем сначала, что для любого μ вычисленные значения последовательности $p_0(\mu)$, $p_1(\mu)$, ..., $p_n(\mu)$ являются точными значениями аналогичной последовательности, соответствующей трехдиагональной матрице, имеющей элементы $\alpha_i + \delta\alpha_i$, $\beta_i + \delta\beta_i$, где $\delta\alpha_i$ и $\delta\beta_i$ удовлетворяют соотношениям

$$\left. \begin{aligned} |\delta\alpha_i| &\leq (3,01) 2^{-t} (|\alpha_i| + |\mu|), \\ |\delta\beta_i| &\leq (1,51) 2^{-t} (|\beta_i|). \end{aligned} \right\} \quad (40.1)$$

Доказательство по индукции. Предположим, что вычисленные

$$p_0(\mu), p_1(\mu), \dots, p_{r-1}(\mu)$$

являются точными для матрицы, имеющей элементы $\alpha_i + \delta\alpha_i$, $\beta_i + \delta\beta_i$, для $i < r$, и покажем, что вычисленное $p_r(\mu)$ является точным значением для матрицы, имеющей те же самые измененные элементы в строках 1, 2, ..., $r-1$ и измененные элементы $\alpha_r + \delta\alpha_r$ и $\beta_r + \delta\beta_r$. Для вычисленного $p_r(\mu)$ имеем

$$\begin{aligned} p_r(\mu) &= fl[(\alpha_r - \mu)p_{r-1}(\mu) - \beta_r^2 p_{r-2}(\mu)] = \\ &= (\alpha_r - \mu)p_{r-1}(\mu)(1 + \varepsilon_1) - \beta_r^2 p_{r-2}(\mu)(1 + \varepsilon_2), \end{aligned} \quad (40.2)$$

где

$$(1 - 2^{-t})^3 \leq 1 + \varepsilon_i \leq (1 + 2^{-t})^3 \quad (i = 1, 2). \quad (40.3)$$

Следовательно, без изменения предыдущих значений $\alpha_i + \delta\alpha_i$ и $\beta_i + \delta\beta_i$ мы можем рассматривать вычисленное $p_r(\mu)$ как точное при

$$\alpha_r + \delta\alpha_r - \mu = (\alpha_r - \mu)(1 + \varepsilon_1), \quad (40.4)$$

$$(\beta_r + \delta\beta_r)^2 = \beta_r^2(1 + \varepsilon_2), \quad (40.5)$$

откуда

$$|\delta\alpha_r| \leq (3,01) 2^{-t} (|\alpha_r| + |\mu|), \quad (40.6)$$

$$|\delta\beta_r| \leq (1,51) 2^{-t} |\beta_r|. \quad (40.7)$$

Теперь наш результат установлен, так как очевидно, что он верен для $r = 1$. Заметим, что (40.7) означает, что если β_r не равно нулю, то $\beta_r + \delta\beta_r$ также не равно нулю и, следовательно, возмущенная матрица имеет различные собственные значения.

Мы можем обозначить возмущенную трехдиагональную матрицу через $(C + \delta C_\mu)$, где индекс μ используется для того, чтобы подчеркнуть, что δC_μ есть функция от μ . Итак, число совпадений знаков между соседними членами вычисленной последовательности равно числу тех нулей матрицы $(C + \delta C_\mu)$, которые больше чем μ .

41. Предположим для удобства, что α_i и β_i удовлетворяют соотношениям

$$|\alpha_i|, |\beta_i| \leq 1. \quad (41.1)$$

Тогда, согласно § 39, все значения μ , которые используются в процессе деления пополам, лежат в интервале $(-3, 3)$. Следовательно,

$$|\delta\alpha_r| < (12,04) 2^{-t}, \quad |\delta\beta_r| \leq (1,51) 2^{-t} \quad (41.2)$$

для любого значения μ . Поэтому матрицы δC_μ равномерно ограничены и, используя ∞ -норму, мы видим, что собственные значения матрицы (δC_μ) всегда лежат в интервалах длины

$$2[12,04 + 2(1,51)] 2^{-t} = (30,12) 2^{-t} = 2d, \quad (41.3)$$

центрами которых являются собственные значения λ_i матрицы C (гл. 2, § 44).

Теперь очень важным вопросом, на который должен быть дан ответ на каждом шаге процесса деления пополам при вычислении k -го собственного значения, является следующий: «Меньше ли, чем k , собственных значений матрицы C , которые больше c_r , средней точки текущего интервала?» Если каждый раз на этот вопрос дается правильный ответ, то λ_k всегда локализуется в точном интервале. Это верно независимо от того, правильно ли определяется действительный номер!

Покажем, что на практике мы должны получить правильный ответ, если c_r находится вне интервала (41.3) с центром в λ_k , как бы близко ни были собственные значения и даже тогда, когда некоторые из других собственных значений матрицы C лежат в λ_k -интервале. Если c_r меньше, чем $\lambda_k - d$, то наверняка первые k собственных значений матрицы $(C + \delta C_\mu)$ должны быть больше, чем c_r , а возможно, что и некоторые другие. Вычисленная последовательность дает правильное число совпадений, соответствующее $(C + \delta C_\mu)$, и, следовательно, на наш вопрос она должна дать правильный ответ «нет» даже тогда, когда получает неправильный действительный номер. Аналогично, если c_r больше, чем $\lambda_k + d$, то наверняка k -е и все меньшие собственные значения матрицы $(C + \delta C_\mu)$ должны быть меньше, чем c_r . Следовательно, мы должны получить на наш вопрос правильный ответ «да».

Итак, в этом процессе возможно следующее: или

(i) правильный ответ получается всегда, и в этом случае λ_k лежит в каждом из интервалов (a_r, b_r) , или

(ii) наступает такой шаг, на котором процесс дает неправильный ответ. Это не может случиться, если только c_r не лежит внутри λ_k -интервала. (Заметим, что ранее здесь могли быть правильные решения, соответствующие c_r , находящимся внутри λ_k -интервала.) Мы покажем, что в этом случае все последовательные интервалы (a_i, b_i) имеют или одну, или обе свои границы внутри λ_k -интервала.

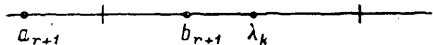


Рис. 3.

По предположению, это верно для (a_{r+1}, b_{r+1}) . Далее, если обе границы лежат внутри λ_k -интервала, то очевидно, что все последующие a_i и b_i также лежат внутри. Если только одна граница находится внутри, например b_{r+1} , то b_{r+1} должно быть точкой, в которой дан неправильный ответ. Итак, ситуация такова, как показано на рис. 3.

Если c_{r+2} , средняя точка (a_{r+1}, b_{r+1}) , находится внутри интервала, то при любом решении c_{r+2} является границей (a_{r+2}, b_{r+2}) . Если c_{r+2} вне интервала, то в отношении c_{r+2} дается правильное решение и мы имеем $b_{r+2} = b_{r+1}$. Следовательно, в любом случае по крайней мере одна из границ (a_{r+2}, b_{r+2}) находится внутри интервала. Продолжая доказательство, получаем наш результат.

Итак, если мы останавливаемся на p -м шаге, то центр c_{p+1} последнего интервала (a_p, b_p) таков, что

$$|c_{p+1} - \lambda_k| < (15,06) 2^{-t} + (b_0 - a_0) 2^{-p-1} < (15,06) 2^{-t} + 3 \cdot 2^{-p}. \quad (41.4)$$

Численный пример

42. В табл. 6 мы показываем применение свойства последовательности Штурма для локализации третьего собственного значения трехдиагональной матрицы пятого порядка, используя 4-значную десятичную арифметику с плавающей запятой. Для этой матрицы ∞ -норма меньше, чем

Таблица 6

i	0	1	2	
α_i	—	0,2165	0,3123	
β_i^2	—	—	0,7163	
$p_i(0,0000)$	+1,0000	+10 ⁰ (0,2165)	—10 ⁰ (0,6487)	
$p_i(0,8500)$	+1,0000	—10 ⁰ (0,6335)	—10 ⁰ (0,3757)	
$p_i(0,4250)$	+1,0000	—10 ⁰ (0,2085)	—10 ⁰ (0,6928)	

i	3	4	5	
α_i	0,4175	0,4431	0,6124	
β_i^2	0,3125	0,2014	0,5016	
$p_i(0,0000)$	—10 ⁰ (0,3385)	—10 ⁻¹ (0,1934)	+10 ⁰ (0,1579)	$s=3$
$p_i(0,8500)$	+10 ⁰ (0,3605)	—10 ⁻¹ (0,7102)	—10 ⁰ (0,1640)	$s=2$
$p_i(0,4250)$	+10 ⁻¹ (0,7035)	+10 ⁰ (0,1408)	—10 ⁻² (0,8902)	$s=2$

1,7, так что все собственные значения лежат в интервале $(-1,7; 1,7)$. Поэтому первая средняя точка есть нуль; в ней получено три совпадения знаков; это показывает, что λ_3 находится в интервале $(0; 1,7)$. Вторая средняя точка есть 0,85, и здесь последовательность Штурма имеет два совпадения, показывая, что λ_3 находится в интервале $(0; 0,85)$. Третья точка есть 0,425 и она дает два совпадения. Следовательно, λ_3 находится между 0 и 0,425. Заметим, что мы получили также следующую информацию: λ_4 и λ_5 лежат между $-1,7$ и 0, λ_1 и λ_2 лежат между 0,85 и 1,7.

Общие замечания о методе деления пополам

43. Если мы реализуем t шагов метода деления пополам, то из (41.4) следует, что оценка ошибки в k -м вычисленном собственном значении равна $(18,06) \cdot 2^{-t}$, и из анализа видно, что, как правило, она будет завышена. Наиболее замечательным в этом результате является то, что оценка не зависит от n .

По существу при каждом вычислении последовательности $p_0, p_1, \dots, p_n$ необходимо выполнить $2n$ умножений и $2n$ вычитаний (β_i^2 могут быть вычислены один раз), так что если независимо находятся все n собственных значений, используя t шагов метода деления пополам, то всего требуется $2n^2t$ умножений и вычитаний. Для больших n время определения собственных значений будет порядка n^2 , тогда как время, необходимое для выполнения как метода Гивенса, так и метода Хаусхолдера,

порядка n^3 . Поэтому для достаточно больших n время вычисления собственных значений мало по сравнению с временем преобразования исходной матрицы к трехдиагональному виду, но из-за множителя $2t$ это неверно, как правило, для матриц меньше 100-го порядка. На практике редко требуются все собственные значения больших матриц.

Время, необходимое для метода деления пополам, не зависит от разделения собственных значений и, как мы показали, на точность не влияет даже их большое скопление. Мы не обязаны использовать метод деления пополам все время и, изолировав один раз собственное значение в некотором интервале, можно подключить итерационный метод, имеющий квадратичную или кубическую сходимость. Мы будем обсуждать такие методы в главе 7. Однако если имеется много близких собственных значений, выигрыш от подключения может быть очень незначительным. Методы, которые имеют квадратичную или кубическую сходимость для простых собственных значений, обычно обладают лишь линейной сходимостью в случае кратных корней. Поэтому, если матрица имеет патологически близкие собственные значения (она не может иметь кратные собственные значения, так как ни одно из β_i не равно нулю), то квадратичная или кубическая сходимость может не наступить и другой метод может быть даже более медленным, чем метод деления пополам.

Метод деления пополам обладает очень большой гибкостью. Мы можем использовать его для нахождения специально выделенных собственных значений $\lambda_{r_1}, \lambda_{r_2}, \dots, \lambda_{r_s}$, собственных значений в заданном интервале (x, y) или для нахождения предписанного числа собственных значений слева или справа от заданного значения. Мы не обязаны определять все требующиеся собственные значения с одной и той же точностью и можем очень просто изменять точность изменением числа шагов деления пополам. Если нас интересует лишь общее распределение собственных значений, а не их точное определение, то метод особенно эффективен. Он был использован Дином (1960) для нахождения распределения собственных значений трехдиагональных матриц до 32 000-го порядка.

Мы обсудили индивидуальное определение каждого собственного значения. Но так как вычисление последовательности Штурма говорит нам и о том, сколько собственных значений лежит справа от точки вычисления, то можно находить все собственные значения одновременно. Если есть несколько точек скопления собственных значений, то выигрыш будет очень ощутимым. Это специальное применение свойства последовательности Штурма было описано Гивенсом (1954) в его оригинальной статье.

Малые собственные значения

44. Заметим, что мы не утверждали, что малые собственные значения будут определяться с малыми относительными ошибками, и простой пример послужит иллюстрацией, почему это не может быть в общем случае. Рассмотрим такую матрицу порядка n :

$$\alpha_i = 2 \quad (i = 1, \dots, n), \quad \beta_{i+1} = 1 \quad (i = 1, \dots, n-1). \quad (44.1)$$

Собственные значения этой матрицы равны $4 \sin^2 [\pi r/2 (n+1)]$ ($r = 1, 2, \dots, n$), и если n велико, то наименьшие собственные значения много меньше единицы. Теперь при вычислении последовательности Штурма для любого значения μ имеем

$$p_r(\mu) = (\alpha_r - \mu) p_{r-1}(\mu) - \beta_r^2 p_{r-2}(\mu), \quad (44.2)$$

и множитель $\alpha_r - \mu$ равен $2 - \mu$. Предположим, например, что мы работаем в 6-значной десятичной арифметике с плавающей запятой и $\mu = 10^{-3}$ (0,213456). Тогда для $2 - \mu$ получаем значение 10^1 (0,199979), причем последние 4 цифры μ полностью не учитываются. Следовательно, вычисленная последовательность Штурма одинакова для всех значений μ между 10^{-3} (0,205) и 10^{-3} (0,215), и мы не можем, вероятно, получить много верных знаков в тех собственных значениях, которые так же малы, как μ . Несмотря на это предупреждение, часто получают малые собственные значения с неожиданно малыми относительными ошибками, но для того, чтобы объяснить эту точность, обычно требуется специальный анализ ошибок.

Бликие собственные значения и малые β_i

45. Мы видели (§ 37), что трехдиагональная матрица обязательно имеет простые собственные значения, если β_i ненулевые. Можно было бы подумать, что матрица, имеющая несколько патологически близких собственных значений, должна иметь по крайней мере один «очень малый элемент» β_i , но это неверно. Можно показать, что если C есть нормированная трехдиагональная матрица и минимум разности между двумя любыми собственными значениями C равен δ , то при фиксированном $n \min \beta_i \rightarrow 0$, когда $\delta \rightarrow 0$, но можно построить матрицы небольшого порядка, имеющие патологически близкие собственные значения и у которых β_i совсем немалые.

В связи с этим мы введем два класса трехдиагональных матриц W_{2n+1}^+ и W_{2n+1}^- , которые также будут представлять интерес и в последующих параграфах. Мы определим W_{2n+1}^+ соотношениями

$$\left. \begin{aligned} \alpha_i &= n+1-i & (i=1, \dots, n+1), \\ \alpha_i &= i-n-1 & (i=n+2, \dots, 2n+1), \\ \beta_i &= 1 & (i=2, \dots, 2n+1), \end{aligned} \right\} \quad (45.1)$$

а W_{2n+1}^- такими соотношениями:

$$\left. \begin{aligned} \alpha_i &= n+1-i & (i=1, \dots, n+1), \\ \alpha_i &= n+1-i & (i=n+2, \dots, 2n+1), \\ \beta_i &= 1 & (i=2, \dots, 2n+1), \end{aligned} \right\} \quad (45.2)$$

так что имеем, например,

$$W_5^+ = \begin{bmatrix} 2 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 0 & 1 & \\ & & 1 & 1 & 1 \\ & & & 1 & 2 \end{bmatrix}, \quad W_5^- = \begin{bmatrix} 2 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 0 & 1 & \\ & & 1 & -1 & 1 \\ & & & 1 & -2 \end{bmatrix}. \quad (45.3)$$

В табл. 7 мы даем собственные значения матриц W_{21}^+ и W_{21}^- с семью десятичными знаками. Матрица W_{21}^- имеет ряд пар собственных значений, которые исключительно близки. В действительности λ_1 и λ_2 совпадают с 15 знаками, λ_3 и λ_4 совпадают с 11 знаками. По сравнению с ними последующие пары постепенно разделяются. Для больших n это явление становится даже более заметным, и можно показать, что $\lambda_1 - \lambda_2$ является приблизительно величиной порядка $(n!)^{-2}$. Матрица W_{21}^+ не нормирована,

Таблица 7

Собственные значения W_{21}^+		Собственные значения W_{21}^-
10,7461942	5,0002444	$\pm 10,7461942$
10,7461942	4,9997825	$\pm 9,2106786$
9,2106786	4,0043540	$\pm 8,0389411$
9,2106786	3,9960482	$\pm 7,0039520$
8,0389411	3,0430993	$\pm 6,0002257$
8,0389411	2,9610589	$\pm 5,0000082$
7,0039522	2,1302092	$\pm 4,0000002$
7,0039518	1,7893214	$\pm 3,0000000$
6,0002340	0,9475344	$\pm 2,0000000$
6,0002175	0,2538058	$\pm 1,0000000$
	-1,1254415	0,0000000

но $\left\| \frac{1}{11} W_{21}^+ \right\|_{\infty} = 1$, так что даже после нормирования мы имеем $\beta_i = 1/11$ и ни одно β_i не может считаться патологически малым. Заметим, что ряд преобладающих собственных значений матрицы W_{21}^- очень близок к некоторым собственным значениям матрицы W_{21}^+ ; с ростом n это становится более заметным. Собственные значения W_{21}^- попарно равны по модулю и противоположны по знаку, и легко показать, что это верно в общем случае для матрицы W_{2n+1}^- .

Собственные векторы матрицы W_{2n+1}^+ могут быть определены через две матрицы U_{n+1} и V_n , имеющие хорошо разделенные собственные значения. Для простоты мы проиллюстрируем это случаем $n = 3$, причем отсюда будет виден и общий случай.

Пусть

$$U_4 = \begin{bmatrix} 3 & 1 & & \\ 1 & 2 & 1 & \\ & 1 & 1 & 1 \\ & & 2 & 0 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 3 & 1 & \\ 1 & 2 & 1 \\ & 1 & 1 \end{bmatrix}. \quad (45.4)$$

Теперь, если U есть собственный вектор матрицы U_4 , соответствующий собственному значению λ , то сразу же видно, что λ есть собственное значение W_7^+ и соответствующий собственный вектор w_1 будет симметричным относительно среднего элемента

$$w_1^T = (u_1, u_2, u_3, u_4, u_3, u_2, u_1). \quad (45.5)$$

С другой стороны, если v — есть собственный вектор матрицы V_3 , соответствующий собственному значению μ , то μ есть собственное значение W_7^+ и соответствующий собственный вектор w_2 будет кососимметричным относительно среднего элемента

$$w_2^T = (v_1, v_2, v_3, 0, -v_3, -v_2, -v_1). \quad (45.6)$$

Поэтому собственные значения матрицы W_{2n+1}^+ те же, что у матриц U_{n+1} и V_n , каждая из которых имеет хорошо разделенные собственные значения. Для $n = 10$ наибольшие собственные значения матриц U_{11} и V_{10} соответственно таковы:

$$10,74619 \ 41829 \ 0339 \dots \text{ и } 10,74619 \ 41829 \ 0332 \dots!$$

Собственные векторы u и v матриц U_{11} и V_{10} приведены в табл. 8. Соответствующие собственные векторы w_1 и w_2 матрицы W_{21}^+ строятся так, как мы только что описали. Заметим, что $(w_1 + w_2)$ почти равен нулю в своей

Т а б л и ц а 8

u	v
1,0000000000	1,0000000000
0,7461941829	0,7461941829
0,3029999415	0,3029999415
0,0859024939	0,0859024939
0,0188074813	0,0188074813
0,0033614646	0,0033614646
0,0005081471	0,0005081471
0,0000665944	0,0000665942
0,0000077063	0,0000077045
0,0000008061	0,0000007905
0,0000001500	

нижней половине, а $(w_1 - w_2)$ — в верхней. Компоненты векторов w_1 и w_2 имеют ярко выраженные минимумы.

Имеет смысл заметить, что всякий раз, когда собственный вектор трехдиагональной матрицы C имеет ярко выраженный минимум во внутренней точке, то C должна иметь несколько очень близких собственных значений. Предположим, например, что нормированная матрица 17-го порядка имеет собственное значение λ , соответствующее собственному вектору u , логарифмы компонент которого приблизительно такие:

$(0, -1, -4, -8, -12, -12, -8, -3, 0, -2, -8, -11, -12, -8, -3, -1, 0)$.

Рассмотрим теперь векторы w_1, w_2, w_3 , где

$$w_1^T = (u_1, u_2, u_3, u_4, 0, \dots, 0), \quad (45.7)$$

$$w_2^T = (0, \dots, 0, u_7, u_8, u_9, u_{10}, u_{11}, 0, \dots, 0), \quad (45.8)$$

$$w_3^T = (0, \dots, 0, u_{14}, u_{15}, u_{16}, u_{17}). \quad (45.9)$$

Все векторы невязки $(Cw_i - \lambda w_i)$ имеют компоненты, которые или равны нулю, или порядка 10^{-11} ; это вытекает сразу же из трехдиагональной природы матрицы C . Следовательно, все три ортогональных вектора w_1, w_2, w_3 дают очень малые невязки, соответствующие значению λ , и поэтому само λ является почти трехкратным собственным значением (гл. 3, § 57).

46. Когда исходная трехдиагональная матрица имеет несколько нулевых β_i , то мы условились рассматривать её как прямую сумму некоторого количества меньших трехдиагональных матриц. Однако предположим, что мы заменили все нулевые β_i на ε (малую величину) и работаем с полученной матрицей $(C + \delta C)$ вместо C . Матрица δC симметричная и все ее собственные значения λ удовлетворяют неравенству

$$|\lambda| < |\delta C|_\infty \leq 2\varepsilon. \quad (46.1)$$

Поэтому все собственные значения матрицы $(C + \delta C)$ различные и лежат в интервалах длины 4ε с центрами в собственных значениях матрицы C . Если компоненты C ограничены по модулю единицей и мы возьмем $\varepsilon = 2^{-l-2}$, то ошибки, полученные от добавления элементов ε , совсем незначительны.

Нам больше не нужно расщеплять $(C + \delta C)$ на меньшие подматрицы, так как она имеет ненулевые внедиагональные элементы. Эта модификация способствует упрощению программы, но несколько неэкономична в смысле времени счета. Однако ее преимущество состоит в том, что C рассматривается как целая, и, следовательно, мы можем находить частное собственное значение, например r -е, не работая отдельно с меньшими матрицами.

Если нас очень интересует время счета, то мы можем сделать шаг в противоположном направлении и заменить все малые β_i в исходной матрице C нулями, получая тем самым матрицу $(C + \delta C)$, которая расщепляется на ряд подматриц. Если мы ликвидируем все β_i , удовлетворяющие неравенству

$$|\beta_i| < \varepsilon, \quad (46.2)$$

то таким же доказательством, какое было проведено выше, можно установить, что максимальная ошибка, сделанная при этом в собственных значениях, равна 2ε . Мы сейчас покажем, что для любого собственного значения, достаточно хорошо отделенного от других, введенная ошибка в действительности много меньше.

Для того чтобы упростить запись, мы предположим, что только β_p и β_q удовлетворяют соотношению (46.2). Поэтому мы работаем с матрицей $(C + \delta C)$, которая является прямой суммой трех трехдиагональных матриц. Обозначим

$$C = \begin{bmatrix} C^{(1)} & & \\ \beta_p & C^{(2)} & \\ & \beta_q & C^{(3)} \end{bmatrix}, \quad C + \delta C = \begin{bmatrix} C^{(1)} & & \\ & C^{(2)} & \\ & & C^{(3)} \end{bmatrix}. \quad (46.3)$$

Предположим, что $C^{(1)}u = \mu u$, так что μ есть собственное значение $C^{(1)}$, а потому и $(C + \delta C)$. Если мы напомним

$$v^T = (u^T | 0, 0, \dots, 0 | 0, 0, \dots, 0), \quad (46.4)$$

где v разделен согласно разделению C , то, определяя η соотношением

$$Cv - \mu v = \eta, \quad (46.5)$$

имеем

$$\eta^T = (0, \dots, 0 | \beta_p v_{p-1}, 0, \dots, 0 | 0, 0, \dots, 0). \quad (46.6)$$

Следовательно, η ортогонален v и μ есть отношение Релея для матрицы C и вектора v (гл. 3, § 54). Предполагая v нормированным, получаем

$$\|\eta\|_2 \leq \varepsilon. \quad (46.7)$$

Отсюда, согласно гл. 3, § 55, вытекает, что если все, кроме одного, собственные значения C отстоят от μ дальше, чем на a , то существует такое собственное значение λ матрицы C , что

$$|\lambda - \mu| < \frac{\varepsilon^2}{a} \left/ \left(1 - \frac{\varepsilon^2}{a^2} \right) \right|. \quad (46.8)$$

Если, например, мы знаем, что минимальное расстояние между собственными значениями равно 10^{-2} , и работаем с 10 десятичными знаками, то можем заменить нулями все β_i , удовлетворяющие неравенству

$$|\beta_i| < 10^{-6}. \quad (46.9)$$

Вычисление собственных значений в арифметике с фиксированной запятой

47. Для метода деления пополам, который мы описали в §§ 37—39, почти необходимо вычисление с плавающей запятой. Если используется фиксированная запятая, то имеет место так много нормировок, что в этом случае вычисление по существу осуществляется с плавающей запятой. Но так как основным требованием является вычисление главных миноров матрицы $(C - \mu I)$ для различных значений μ , мы можем использовать метод, описанный в гл. 4, § 49. Так как $(C - \mu I)$ имеет трехдиагональный вид, находим, что данный метод идентичен методу Гаусса с выбором главного элемента по столбцу. Поскольку этот процесс снова потребуется, когда мы будем рассматривать задачу определения собственных векторов, опишем его подробно.

Процесс состоит из $n - 1$ основных шагов, причем строки $r + 1$, $r + 2, \dots, n$ не изменяются до начала r -го основного шага. Схема расположения элементов перед началом этого шага для $n = 6$, $r = 3$ такова:

$$\begin{matrix} m_2 \\ m_3 \end{matrix} \begin{bmatrix} u_1 & v_1 & w_1 & & & \\ & u_2 & v_2 & w_2 & & \\ & & p_3 & q_3 & & \\ & & \beta_4 & \alpha_4 - \mu & \beta_5 & \\ & & & \beta_5 & \alpha_5 - \mu & \beta_6 \\ & & & & \beta_6 & \alpha_6 - \mu \end{bmatrix}. \quad (47.1)$$

Система обозначений станет ясной, когда мы опишем r -й шаг, состоящий в следующем:

(i) Если $|\beta_{r+1}| > p_r$, то переставляем строки r и $r + 1$. Обозначим элементы полученной строки r через u_r, v_r, w_r и строки $r + 1$ — через $x_{r+1}, y_{r+1}, z_{r+1}$, так что, если имела место перестановка, то

$$u_r = \beta_{r+1}, \quad v_r = \alpha_{r+1} - \mu, \quad w_r = \beta_{r+2}; \quad x_{r+1} = p_r, \quad y_{r+1} = q_r, \quad z_{r+1} = 0,$$

и в противном случае

$$u_r = p_r, \quad v_r = q_r, \quad w_r = 0; \quad x_{r+1} = \beta_{r+1}, \quad y_{r+1} = \alpha_{r+1} - \mu, \quad z_{r+1} = \beta_{r+2}.$$

Схема, соответствующая (47.1), теперь будет такой:

$$\begin{matrix} m_2 \\ m_3 \end{matrix} \begin{bmatrix} u_1 & v_1 & w_1 & & & \\ & u_2 & v_2 & w_2 & & \\ & & u_3 & v_3 & w_3 & \\ & & & x_4 & y_4 & z_4 \\ & & & & \beta_5 & \alpha_5 - \mu & \beta_6 \\ & & & & & \beta_6 & \alpha_6 - \mu \end{bmatrix}, \quad (47.2)$$

где или w_3 , или z_4 равно нулю.

(ii) Вычисляем $m_{r+1} = x_{r+1}/u_r$ и сохраняем его. Заменяем x_{r+1} нулем

(iii) Вычисляем $p_{r+1} = y_{r+1} - m_{r+1}v_r$, $q_{r+1} = z_{r+1} - m_{r+1}w_r$ и записываем их соответственно на места y_{r+1} и z_{r+1} . Теперь схема такова:

$$\begin{matrix} & & & & & & \\ & & & & & & \\ m_2 & & & & & & \\ m_3 & & & & & & \\ m_4 & & & & & & \end{matrix} \begin{bmatrix} u_1 & v_1 & w_1 & & & \\ & u_2 & v_2 & w_2 & & \\ & & u_3 & v_3 & w_3 & \\ & & & p_4 & q_4 & \\ & & & \beta_5 & \alpha_5 - \mu & \beta_6 \\ & & & & \beta_6 & \alpha_6 - \mu \end{bmatrix}. \quad (47.3)$$

Заметим, что $p_1 = \alpha_1 - \mu$, $q_1 = \beta_2$. Простое доказательство по индукции показывает, что если C нормирована так, что

$$|\alpha_i|, |\beta_i| < \frac{1}{5},$$

то для любого допустимого значения μ все элементы, полученные во время преобразования, ограничены единицей, так что вычисление с фиксированной запятой вполне удобно. Для того чтобы подчеркнуть, что этот процесс действительно есть метод Гаусса, мы показали $(C - \mu I)$ и последовательно преобразованные матрицы полностью, но на практике будем, конечно, использовать преимущества трехдиагональной формы и хранить только u_i, v_i, w_i как линейные массивы. Если мы сохраним m_i, u_i, v_i, w_i и запомним, имела ли место перестановка перед вычислением m_i , то сможем впоследствии обрабатывать правую часть уравнения

$$(C - \mu I)x = b.$$

r -ый главный минор d_r равен $(-1)^{k_r} u_1 u_2 \dots u_{r-1} p_r$, где k_r есть общее число перестановок, которые имели место до конца $(r - 1)$ -го основного шага. Конечно, нам не нужно вычислять это произведение; мы просто запоминаем его знак. Так как

$$\begin{aligned} d_{r+1}/d_r &= (-1)^{k_{r+1}} u_1 u_2 \dots u_r p_{r+1} / (-1)^{k_r} u_1 u_2 \dots u_{r-1} p_r = \\ &= (-1)^{k_{r+1} - k_r} u_r p_{r+1} / p_r, \end{aligned} \quad (47.4)$$

то видим, что d_{r+1} и d_r имеют один и тот же знак, если правая часть (47.4) положительна.

Невозможно выполнить анализ ошибок, подобный тому, который мы дали в § 41. В результате перестановок эквивалентное возмущение исходной матрицы может быть несимметричным. Если, например, перестановки имеют место на каждом этапе, то матрица, которая точно дает вычисленные множители и ведущие строки, имеет такой вид:

$$\begin{bmatrix} \alpha_1 + \varepsilon_1 & \beta_2 + \varepsilon_2 & \varepsilon_3 & \varepsilon_4 & \varepsilon_5 \\ & \beta_2 & \alpha_2 & \beta_3 & \\ & & \beta_3 & \alpha_3 & \beta_4 \\ & & & \beta_4 & \alpha_4 & \beta_5 \\ & & & & \beta_5 & \alpha_5 \end{bmatrix}, \quad (47.5)$$

причем все возмущения находятся в первой строке. Далее, эквивалентное возмущение, соответствующее r -му шагу, изменяет некоторые из эквивалентных возмущений, соответствующих более ранним шагам.

Так как возмущения несимметричны, то некоторые матрицы главных миноров могут иметь даже комплексно сопряженные нули.

Однако легко показать, что 2-нормы всех матриц возмущения ограничены величиной $Kn^{1/2}2^{-l}$ для некоторого K и, следовательно, знаки всех главных миноров должны быть получены правильно, за исключением, возможно, знаков для тех значений μ , которые лежат в узких интервалах с центрами в $(1/2)n(n+1)$ собственных значениях всех матриц главных миноров. Следовательно, не так уж неожиданно, что процесс деления пополам, основанный на этом методе, весьма эффективен. На практике мы установили, что для матрицы, нормированной так, как мы описали, ошибка в вычисленном собственном значении редко достигает $2 \cdot 2^{-l}$ даже для матриц с патологически близкими собственными значениями. Есть большое желание использовать этот процесс вместо процесса §§ 37—39, так как метод, который мы будем рекомендовать для вычисления собственных векторов, также требует триангуляризации матрицы $(C - \mu I)$. На вычислительных машинах, у которых количество разрядов мантиссы числа с плавающей запятой значительно меньше количества разрядов числа с фиксированной запятой, только что описанный нами метод почти наверняка будет давать более точные результаты. Все, чего нам недостает, это строгой *априорной* оценки ошибок!

Вычисление собственных векторов трехдиагональной матрицы

48. Теперь обратимся к вычислению собственных векторов симметричной трехдиагональной матрицы C . Опять предположим, что ни один элемент β_i не равен нулю, так как в противном случае, как указано в § 36, проблема определения собственных векторов матрицы C сводится к аналогичной проблеме для ряда меньших трехдиагональных матриц. Предположим, что очень точные собственные значения уже получены по методу деления пополам или по некоторому другому методу сравнимой точности.

На первый взгляд проблема кажется очень простой. Мы знаем, что компоненты собственного вектора x , соответствующего λ , удовлетворяют уравнениям

$$(\alpha_1 - \lambda)x_1 + \beta_2x_2 = 0, \quad (48.1)$$

$$\beta_ix_{i-1} + (\alpha_i - \lambda)x_i + \beta_{i+1}x_{i+1} = 0 \quad (i = 2, \dots, n-1), \quad (48.2)$$

$$\beta_nx_{n-1} + (\alpha_n - \lambda)x_n = 0. \quad (48.3)$$

Очевидно, что x_1 не может быть нулем, так как из уравнений (48.1) и (48.2) следует, что и все другие x_i должны тогда равняться нулю. Так как нас не интересует нормирующий множитель, то мы можем взять $x_1 = 1$. Простое доказательство по индукции показывает, что

$$x_r = (-1)^{r-1} p_{r-1}(\lambda) / \beta_2 \beta_3 \dots \beta_r \quad (r = 2, \dots, n). \quad (48.4)$$

Для того чтобы получить этот результат, нам нужны только уравнения (48.1) и (48.2), но полученные x_{n-1} и x_n автоматически удовлетворяют (48.3), так как мы имеем

$$\begin{aligned} \beta_nx_{n-1} + (\alpha_n - \lambda)x_n &= (-1)^{n-1} \left[\frac{(\alpha_n - \lambda)p_{n-1}(\lambda)}{\beta_2\beta_3 \dots \beta_n} - \frac{\beta_n p_{n-2}(\lambda)}{\beta_2\beta_3 \dots \beta_{n-1}} \right] = \\ &= (-1)^{n-1} p_n(\lambda) / \beta_2\beta_3 \dots \beta_n = 0; \end{aligned} \quad (48.5)$$

последнее равенство справедливо, потому что $p_n(\lambda)$ равно нулю, когда λ есть собственное значение.

Итак, мы имеем явные выражения для компонент собственного вектора через $p_r(\lambda)$. Мы видели, что $p_r(\lambda)$ определяли собственные значения исключительно точно, и поэтому могли бы предположить, что если λ было очень хорошим приближением к собственному значению, то использование этих явных выражений для компонент даст хорошее приближение к собственному вектору. В действительности это не так. Покажем, что вектор, полученный таким путем при хорошем приближении к собственному значению, может содержать катастрофически большую ошибку.

Неустойчивость явного выражения для собственного вектора

49. Пусть λ очень близко к собственному значению, но не точно равно ему, и пусть x_r — точные компоненты, полученные из соотношений (48.4), соответствующих этому значению λ . Тогда по определению x_r удовлетворяют уравнениям (48.1) и (48.2). Компоненты x_{n-1} и x_n не могут удовлетворять уравнению (48.3) точно, потому что в противном случае x был бы ненулевым вектором, удовлетворяющим уравнению $(A - \lambda I)x = 0$, и следовательно, λ должно быть точным собственным значением, вопреки нашему предположению. Поэтому имеем

$$\beta_n x_{n-1} + (\alpha_n - \lambda) x_n = \delta \neq 0, \quad (49.1)$$

и вектор x точно удовлетворяет уравнению

$$(C - \lambda I)x = \delta e_n, \quad (49.2)$$

где e_n есть n -й столбец единичной матрицы. Так как нас не интересует произвольный ненулевой множитель, то можно считать, что x есть решение уравнения

$$(C - \lambda I)x = e_n. \quad (49.3)$$

Матрица $(C - \lambda I)$, по предположению, невырожденная и, следовательно,

$$x = (C - \lambda I)^{-1} e_n. \quad (49.4)$$

Теперь, для того чтобы подчеркнуть, что явление, которое мы должны продемонстрировать, не связано с плохой обусловленностью проблемы собственных векторов для C , предположим, что C имеет хорошо разделенные собственные значения, так что в силу симметрии C собственные векторы не чувствительны к возмущениям в ее элементах. Предположим, что собственные значения упорядочены так, что

$$\lambda_1 > \lambda_2 > \dots > \lambda_n, \quad (49.5)$$

и что λ очень близко к λ_k . Следовательно, $\lambda - \lambda_k$ «мало», но $\lambda - \lambda_i$ ($i \neq k$) «не малы».

Пусть $u_1, u_2, \dots, u_n$ — ортонормальная система собственных векторов C . Тогда мы можем представить e_n в виде

$$e_n = \sum_1^n \gamma_i u_i, \quad \sum_1^n \gamma_i^2 = 1. \quad (49.6)$$

Поэтому уравнение (49.4) дает

$$x = \sum_1^n \gamma_i u_i / (\lambda_i - \lambda) = \gamma_k u_k / (\lambda_k - \lambda) + \sum_{i \neq k} \gamma_i u_i / (\lambda_i - \lambda). \quad (49.7)$$

Если вектор x должен быть хорошим приближением к u_k , то важно, чтобы $\gamma_k/(\lambda_k - \lambda)$ было велико по сравнению с $\gamma_i/(\lambda_i - \lambda)$ ($i \neq k$). Мы знаем, что $\lambda_k - \lambda$ очень мало, но если γ_k также очень мало, то x будет далек от u_k . Предположим, например, что мы работаем с десятью десятичными знаками в $\lambda_k - \lambda = 10^{-10}$. Если оказалось, что γ_k равно 10^{-18} , то x будет почти ортогонален вектору u_k .

Читатель может почувствовать, что это доказательство довольно экстравагантно и что нет причин ожидать, что вектор e_n будет иметь столь малую компоненту в направлении какого-либо собственного вектора. Утверждаем, что это явление достаточно общее, и основная цель следующих нескольких параграфов заключается в том, чтобы показать, насколько легко оно может произойти.

50. Из уравнения (49.6) и ортогональности u_i имеем

$$\gamma_k = u_k^T e_n = n\text{-й компоненте } u_k. \quad (50.1)$$

Поэтому, если последняя компонента нормированного вектора u_k очень мала, то собственный вектор, полученный из явных выражений (48.4), соответствующих очень хорошему приближению λ к λ_k , не будет хорошим приближением к u_k .

До рассмотрения вопроса о том, из-за каких факторов n -я компонента u_k может стать патологически малой, вернемся к системам уравнений (48.1), (48.2), (48.3). Вектор x , определенный соотношением (48.4), есть точное решение первых $(n - 1)$ уравнений системы

$$(C - \lambda I)x = 0. \quad (50.2)$$

Так как λ не является точным собственным значением, то лишь нулевой вектор будет решением полной системы. Вместо того чтобы рассматривать решения первых $(n - 1)$ из этих уравнений, рассмотрим систему, полученную из данной исключением r -го уравнения. Мы снова будем иметь ненулевое решение, которое удовлетворяет всем уравнениям (50.2), за исключением r -го. Следовательно, этот вектор x есть точное решение системы

$$(C - \lambda I)x = \delta e_r \quad (\delta \neq 0). \quad (50.3)$$

Точно таким же доказательством, какое было дано выше, мы устанавливаем, что x будет хорошим приближением к u_k при условии, что r -я компонента u_k не мала. Итак, если наибольшая компонента u_k есть r -я, то при вычислении u_k следует исключить r -е уравнение. Этот результат поучительный, но практически не очень полезен, так как мы не знаем априори положение наибольшей компоненты u_k . В действительности u_k есть именно тот вектор, который мы пытаемся вычислить!

Решение, соответствующее исключению r -го уравнения, может быть получено следующим образом. Берем $x_1 = 1$ и вычисляем $x_2, x_3, \dots, x_r$, используя уравнение (48.1) и уравнения (48.2) до $i = r - 1$. Затем берем $x'_n = 1$ и вычисляем $x'_{n-1}, x'_{n-2}, \dots, x'_r$, используя уравнение (48.3) и уравнения (48.2) для $i = n - 1, n - 2, \dots, r + 1$. Комбинируем эти два частичных вектора, выбирая k так, чтобы

$$x_r = kx'_r, \quad (50.4)$$

и берем

$$\hat{x}^T = (x_1, x_2, \dots, x_r, kx'_{r+1}, kx'_{r+2}, \dots, kx'_n). \quad (50.5)$$

Если этот процесс выполняется без ошибок округления, то очевидно, что мы имеем точное решение (50.3) для некоторого ненулевого δ .

Численные примеры

51. Рассмотрим теперь тривиальный численный пример для матрицы второго порядка. Пусть матрица C определена так:

$$C = \begin{bmatrix} 0,713263 & 0,000984 \\ 0,000984 & 0,121665 \end{bmatrix}, \quad \left. \begin{array}{l} \lambda_1 = 0,71326463 \dots \\ \lambda_2 = 0,12166336 \dots \end{array} \right\}, \quad (51.1)$$

и возьмем приближение $\lambda = 0,713265$ к λ_1 ; оно имеет ошибку 10^{-7} (0,36...). Рассмотрим собственный вектор, полученный исключением второго уравнения системы $(C - \lambda I)x = 0$. Мы имеем

$$-0,000002x_1 + 0,000984x_2 = 0, \quad (51.2)$$

откуда

$$x^T = (1, \quad 0,002033). \quad (51.3)$$

Если исключим первое уравнение, то имеем

$$0,000984x_1 - 0,591600x_2 = 0, \quad (51.4)$$

откуда

$$x^T = (1, \quad 0,00166329). \quad (51.5)$$

Рассматривая, что могло бы быть, если бы мы использовали точное значение λ_1 , видим, что второй вектор точнее. Первый вектор имеет только три верных знака, тогда как второй — почти восемь. Это предсказывалось нашим анализом предыдущего параграфа. Наибольшей компонентой u_1 является первая, поэтому нам следует исключить первое уравнение. Возвращаясь к вычислению собственного вектора, соответствующего λ_2 , находим, что теперь должно быть исключено второе уравнение. Заметим, что задача определения этих собственных значений и собственных векторов хорошо обусловлена. Поэтому трудности, которые здесь возникают, не могут быть отнесены к плохой обусловленности решаемой задачи.

52. Приведенный пример, конечно, был несколько искусственным, но это неизбежно для матриц столь малого порядка. В качестве второго примера рассмотрим матрицу W_{21}^- из § 45. Она имеет хорошо разделенные собственные значения, и, следовательно, в силу симметрии задача определения собственных векторов хорошо обусловлена. Рассмотрим собственный вектор, соответствующий наибольшему собственному значению λ_1 . Если мы имеем точное λ_1 , то собственный вектор может быть точно получен с помощью решения любых 20 из 21 уравнений

$$(W_{21}^- - \lambda_1 I)x = 0. \quad (52.1)$$

Предположим, что мы используем последние 20 уравнений, и возьмем $x_{21} = 1$. Тогда

$$x_{20} - (10 + \lambda_1) = 0, \quad (52.2)$$

$$x_s = (\lambda_1 + s - 10)x_{s+1} - x_{s+2} \quad (s = 19, 18, \dots, 1). \quad (52.3)$$

Далее, λ_1 больше 10, и, следовательно, если мы предположим, что $x_{s+1} > x_{s+2} > 0$, то

$$x_s > sx_{s+1} - x_{s+2} > (s-1)x_{s+1}. \quad (52.4)$$

Таблица 9

u_1	x	y	z	w
1,0000000	1,0000000	1,0000000	1,0000000	$1,414404161 \times 10^{19}$
0,74619418	0,74619420	0,74619420	0,74619418	$1,055420141 \times 10^{19}$
0,30299994	0,30299998	0,30299998	0,30299993	$4,285643680 \times 10^{18}$
0,08590249	0,08590260	0,08590259	0,08590248	$1,215008414 \times 10^{18}$
0,01880748	0,01880783	0,01880780	0,01880742	$2,66013790 \times 10^{17}$
0,00336146	0,00336303	0,00336288	0,00336120	$4,75446937 \times 10^{16}$
0,00050815	0,00051681	0,00051596	0,00050668	$7,18725327 \times 10^{15}$
0,00006659	0,00012346	0,00011789	0,00005694	$9,41912640 \times 10^{14}$
0,00000771	0,0004395.	0,00039724	-0,00006562	$1,089849597 \times 10^{14}$
0,00000080	0,003721..	0,00335645	-0,0006309.	$1,129098198 \times 10^{13}$
0,00000007	0,03582...	0,03231537	-0,006083..	$1,059143420 \times 10^{12}$
0,00000001	0,3812...	0,34391079	-0,066001..	$9,07788963 \times 10^{10}$
0,00000000	$0,4442 \times 10^4$	4,00732756	-0,76917...	$7,16312559 \times 10^9$
0,00000000	$0,5624 \times 10^2$	50,73426451	-0,738...	$5,23693576 \times 10^8$
0,00000000	$0,7686 \times 10^3$	Здесь компоненты увеличиваются так же быстро, как компоненты x .		$3,56680106 \times 10^7$
0,00000000	$0,4128 \times 10^5$			$2,27385533 \times 10^6$
0,00000000	$0,1768 \times 10^6$			$1,36242037 \times 10^5$
0,00000000	$0,2950 \times 10^7$			$7,70028174 \times 10^3$
0,00000000	$0,5217 \times 10^8$			$4,08658380 \times 10^2$
0,00000000	$0,9751 \times 10^9$			$2,07461942 \times 10^1$
0,00000000	$0,1920 \times 10^{11}$			$1,00000000 \times 10^0$

Так как очевидно, что x_{20} больше, чем x_{21} , то это показывает, что x_i растут с уменьшением i по крайней мере до x_2 и

$$x_2 > x_3 > 2! \quad x_4 > 3! \quad x_5 > \dots > 19! \quad x_{21} = 19! > 10^{17}. \quad (52.5)$$

В действительности x_1 также больше, чем x_2 , так как

$$(\lambda_1 - 10)x_1 = x_2, \quad (52.6)$$

а λ_1 лежит между 10 и 11. Следовательно, последняя компонента *нормированного* собственного вектора меньше, чем 10^{-17} , и поэтому в обозначениях § 49 мы имеем $\gamma_1 < 10^{-17}$. Поэтому явное решение (48.4), соответствующее приближенному λ , имеющему ошибку лишь 10^{-10} , почти ортогонально u_1 .

Это иллюстрируется табл. 9. Здесь u_1 есть точный собственный вектор, заданный с 8 знаками после запятой, а x — вектор, полученный точным решением первых 20 уравнений $(A - \lambda I)x = 0$ с $\lambda = 10,74619420$. Так как этот вектор трудно вычислять, мы не дали всюду одно и то же число знаков. Однако все приведенные знаки верные и они показывают поведение x . *Заметим, что x есть вектор, соответствующий точному применению явных формул (48.4).*

Вектор y есть вектор, который мы действительно получили из первых 20 уравнений при том же значении λ , используя 8-значное вычисление с фиксированной запятой. Вследствие ошибок округления (и только их) y отличается от x , хотя в целом он ведет себя так же в том смысле, что его компоненты быстро уменьшаются до минимума и затем снова возрастают. Очевидно, что это неустойчивый процесс; ошибки округления, сделанные в более ранних компонентах, катастрофически увеличиваются в последних. Это может привести к заключению, что отклонение y от точного собственного вектора u_1 есть результат действия именно этих ошибок округления и что при определенной точности λ должен бы получиться хороший собственный вектор, если мы решали первые 20 уравнений, используя арифметику повышенной точности. Для того чтобы показать ошибочность этого предположения, и дан вектор x .

Четвертый вектор z в табл. 9 есть точное решение первых 20 уравнений с $\lambda = 10,74619418$, и снова все приведенные знаки верные. Подобно вектору x компоненты сначала быстро уменьшаются, а затем увеличиваются, хотя последние компоненты здесь отрицательны. *Из x и z мы видим, что нет десятизначного значения λ , для которого точное решение первых 20 уравнений дает что-либо похожее на u_1 !*

Наконец, мы даем вектор w , полученный решением последних 20 уравнений. Как показал наш анализ, компоненты быстро уменьшаются. После нормирования этот вектор совпадает с u до единицы восьмого знака.

Обратная итерация

53. Рассмотрим теперь решение неоднородной системы уравнений

$$(A - \lambda I)x = b, \quad (53.1)$$

где b будем считать на время произвольным нормированным вектором. Если b представлено в виде

$$b = \sum_{i=1}^n \gamma_i u_i, \quad (53.2)$$

то точное решение (53.1) таково:

$$x = \sum_{i=1}^n \gamma_i u_i / (\lambda_i - \lambda). \quad (53.3)$$

Предполагая на некоторое время, что мы можем решать такие системы уравнений, как (53.1), без ошибок округления, мы видим, что если λ близко к λ_k , но не близко к другим λ_i , то разложение x содержит значительно большую часть вектора u_k , чем разложение b . Для того чтобы x было хорошим приближением к u_k , необходимо, чтобы γ_k не было малым. Если теперь мы решаем систему уравнений

$$(A - \lambda I) y = x \quad (53.4)$$

без ошибок округления, то получим

$$y = \sum_{i=1}^n \gamma_i u_i / (\lambda_i - \lambda)^2, \quad (53.5)$$

и разложение y содержит еще большую часть вектора u_k , чем разложение x .

Очевидно, что мы могли бы бесконечно повторять этот процесс, получая тем самым такие векторы, которые были бы все более и более близки к u_k , но при условии, что у нас есть некоторый метод выбора b , не приводящий к патологической ортогональности вектору u_k ; уже второй вектор y обычно будет очень хорошим приближением к u_k . Предположим, например, что

$$\lambda - \lambda_k = 10^{-10}, \quad |\lambda - \lambda_i| > 10^{-3} \quad (i \neq k), \quad \gamma_k = 10^{-4}. \quad (53.6)$$

В этом случае мы могли бы с уверенностью сказать, что разложение вектора b содержит малую часть вектора u_k и что λ_k не очень хорошо отделено от других собственных значений.

Все же мы имеем

$$y = \gamma_k u_k / (\lambda_k - \lambda)^2 + \sum_{i \neq k} \gamma_i u_i / (\lambda_i - \lambda)^2 = 10^{16} u_k + \sum_{i \neq k} \gamma_i u_i / (\lambda_i - \lambda)^2, \\ 10^{-16} y = u_k + 10^{-16} \sum_{i \neq k} \gamma_i u_i / (\lambda_i - \lambda)^2, \quad (53.7)$$

и

$$\| 10^{-16} \sum_{i \neq k} \gamma_i u_i / (\lambda_i - \lambda)^2 \|_2 < 10^{-16} \cdot 10^6 \left[\sum_{i \neq k} \gamma_i^2 \right]^{1/2} \leq 10^{-10}. \quad (53.8)$$

Итак, даже в таком неблагоприятном случае нормированный вектор y почти точно совпадает с u_k .

Выбор начального вектора b

54. Остается задача о выборе вектора b . Заметим, что мы не требуем, чтобы b был выбран особенно хорошо, т. е. чтобы компонента с u_k была одной из самых больших, мы требуем лишь, чтобы доля u_k в разложении вектора b не была слишком малой. Не найдено ни одного такого решения этой задачи, для которого можно было бы строго доказать его удовлетворительность во всех случаях, но следующий метод оказался на практике

достаточно эффективным, и есть веские основания верить в то, что он не может оказаться несостоятельным.

Соответственно хорошему приближению λ выполняется исключение Гаусса с перестановками для матрицы $(C - \lambda I)$ так, как было описано в § 47. (Сейчас безразлично, выполняются ли вычисления с фиксированной или плавающей запятой.) Пусть сохраняются ведущие строки u_i , v_i , w_i и множители m_i вместе с информацией о перестановках. Поэтому наша последняя запоминаемая матрица U будет иметь такой же вид, как для случая $n = 5$:

$$U = \begin{bmatrix} u_1 & v_1 & w_1 & & \\ & u_2 & v_2 & w_2 & \\ & & u_3 & v_3 & w_3 \\ & & & u_4 & v_4 \\ & & & & p_5 \end{bmatrix}. \quad (54.1)$$

Заметим, что ни одно из u_i не может быть равно нулю, так как на каждом этапе преобразования ведущий элемент выбирается из β_{r+1} и p_r , а мы все еще предполагаем, что β_i не равны нулю. Последний элемент p_n , конечно, может быть нулем, и в действительности он бы им и был, если бы λ было точным собственным значением и точно выполнялись вычисления (см., однако, § 56).

Теперь выберем правую часть b такой, чтобы при выполнении процесса исключения она давала в качестве преобразованной правой части вектор e , состоящий целиком из элементов, равных единице. Нам не нужно определять сам вектор b . Заметим, что b есть функция от λ , так что мы эффективно используем для каждого λ различные правые части. Если теперь решаем систему

$$Ux = e \quad (54.2)$$

с помощью обратной подстановки, то тем самым выполняем один шаг обратной итерации. Если p_n равно нулю, то мы можем заменить его на подходящее малое число, например $2^{-t} \|C\|_\infty$. Теперь может быть сделано столько шагов обратной итерации, сколько требуется, используя множители m_i для выполнения преобразования и матрицу U для обратной подстановки. До тех пор, пока мы не захотим вычислить следующий собственный вектор, относящийся к другому λ , не потребуется никаких дальнейших триангуляризаций.

На практике мы обычно убеждались, что такой выбор b настолько эффективен, что уже после первой итерации x было *хорошим* приближением к требуемому собственному вектору, а третья итерация реально *никогда* не была существенной.

Анализ ошибок

55. В главе 9 мы обсудим обратную итерацию в более общем плане; кроме того, случай симметричной трехдиагональной матрицы уже был подробно изучен Уилкинсоном (1963b, стр. 143—147). По этой причине мы довольствуемся здесь менее формальным исследованием.

Заметим сначала, что для данного приближенного собственного значения мы всегда используем одно и то же треугольное разложение матрицы $(C - \lambda I)$, и оно будет точным треугольным разложением $(C - \lambda I + F)$ для некоторой малой матрицы ошибок F . Отсюда следует, что

в лучшем случае мы можем получить, используя возможно большее число итераций, только собственный вектор матрицы $(C + F)$. Мы не можем ожидать, что получим правильно все знаки в собственном векторе матрицы C . Но это верно, вообще говоря, для любого метода, и если C предварительно получена по методу Гивенса или Хаусхолдера, то ошибки этого вида будут уже введены.

Предположим для удобства, что C нормирована, и выполним такие итерации:

$$(C - \lambda I) x_{r+1} = y_r, \quad y_{r+1} = x_{r+1} / \|x_{r+1}\|_\infty, \quad (55.1)$$

так что x_r нормируется после каждой итерации. Теперь мы хотим решить вопрос, когда же на практике следует оканчивать итерации. Используя преимущества трехдиагональной природы матрицы C , мы можем применить доказательство гл. 4, § 63, для того чтобы показать, что вычисленное решение удовлетворяет уравнению

$$(C - \lambda I + G) x_{r+1} = y_r, \quad (55.2)$$

где

$$\|G\|_2 \leq kn^{1/2} 2^{-t}. \quad (55.3)$$

Если обозначим

$$\lambda = \lambda_k + \varepsilon, \quad (55.4)$$

то (55.2) может быть записано в виде

$$(C - \lambda_k I) y_{r+1} = \varepsilon y_{r+1} - G y_{r+1} + y_r / \|x_{r+1}\|_\infty. \quad (55.5)$$

Из членов справа первый мал, так как λ по предположению близко к собственному значению, второй мал согласно (55.3). Третий будет мал, если $\|x_{r+1}\|_\infty$ велика. Следовательно, чем больше мы получим значение для $\|x_{r+1}\|_\infty$, тем меньше будет наша оценка для правой части (55.5), и это есть вектор-невязка, соответствующий y_{r+1} и λ_k . Уилкинсоном (1963b, стр. 139—147), показано, что $\|x_{r+1}\|_\infty$ должна быть большой, если разложение y_r содержит не очень малую часть вектора u_k .

Отсюда мы получили следующее эмпирическое правило для вычислений с плавающей запятой на АСЕ с $t = 46$: *продолжаем итерации до тех пор, пока не выполнится условие*

$$\|x_r\|_\infty \geq 2^t / 100n, \quad (55.6)$$

и затем выполняем еще один шаг.

На практике это никогда не приводило более чем к трем итерациям, и обычно было достаточно двух. (Заметим, что первая итерация включает лишь обратную подстановку.) Множитель $1/100n$ не имеет глубокого смысла, а только обеспечивает то, что мы будем редко выполнять необязательную дополнительную итерацию. Можно было бы думать, что нормированный вектор y_r должен стремиться к пределу в рамках ошибки округления и что некоторое условие, основанное на этом, должно быть более естественным критерием окончания, чем (55.6). Однако это не так; если C имеет более чем одно собственное значение, патологически близкое к λ , то в конечном счете y_r будет лежать в подпространстве, натянутом на соответствующие собственные векторы. Иногда случается, что при последовательных итерациях мы получаем совсем разные векторы, лежащие в этом подпространстве. Тогда последовательные y_r не «сходятся» совсем.

Численный пример

56. В табл. 10 показано применение этого метода к вычислению доминирующего собственного вектора матрицы W_{21}^- . Мы взяли $\lambda = 10,7461942$ (верное с 9 знаками) и показываем матрицу ведущих строк.

Таблица 10

Нормированный				
u	v	w	x	x
1,00000 00	-4,74619 42	1,00000 00	-2,35922 2×10^7	1,00000 00
1,00000 00	-2,74619 42	1,00000 00	-1,76043 8×10^7	0,74619 43
1,00000 00	-3,74619 42	1,00000 00	-7,14844 2×10^6	0,30300 00
1,00000 00	-4,74619 42	1,00000 00	-2,02663 1×10^6	0,08590 25
1,00000 00	-5,74619 42	1,00000 00	-4,43710 5×10^5	0,01880 75
1,00000 00	-6,74619 42	1,00000 00	-7,93046 7×10^4	0,00336 15
1,00000 00	-7,74619 42	1,00000 00	-1,19885 3×10^4	0,00050 82
1,00000 00	-8,74619 42	1,00000 00	-1,57128 3×10^3	0,00006 66
1,00000 00	-9,74619 42	1,00000 00	-1,81935 8×10^2	0,00000 77
1,00000 00	-10,74619 42	1,00000 00	-1,89629 9×10	0,00000 08
1,00000 00	-11,74619 42	1,00000 00	-1,88118 0	0,00000 01
-4,07784 23	0,34996 24	Перестановки	-0,25253 85	0,00000 00
-12,66037 37	1,00000 00	прекратились	-0,08518 65	0,00000 00
-13,66720 76	1,00000 00	с 12-го шага	-0,07849 27	0,00000 00
-14,67302 64	1,00000 00		-0,07277 54	0,00000 00
-15,67804 19	1,00000 00		-0,06783 52	0,00000 00
-16,68241 07	1,00000 00		-0,06352 36	0,00000 00
-17,68625 08	1,00000 00		-0,05972 74	0,00000 00
-18,68965 31	1,00000 00		-0,05635 39	0,00000 00
-19,69268 87	1,00000 00		-0,05323 40	0,00000 00
-20,69541 39 = p_{21}			-0,04831 99	0,00000 00

Для облегчения изложения мы не даем эту матрицу в форме (54.1), но выписали u_i , v_i и w_i по столбцам. Перестановки имели место на каждом из первых 11 этапов, поэтому первые 11 ведущих строк имеют три ненулевых элемента. Далее перестановок не было. Заметим, что столбец u дает ведущие элементы, и ни один из них совсем не мал (ср. гл. 4, § 66). Точное вычисление для *точного* собственного значения должно бы дать нулевое значение для p_n , последнего ведущего элемента. Однако, даже если бы было выполнено точное вычисление с $\lambda = 10,7461942$, ведущие элементы должны были бы быть почти точно такими, как показаны (первые 11 элементов по-прежнему должны быть равны единице), за исключением двенадцатого, который может полностью отличаться. Мы должны иметь значение λ значительно более точное, чем то, которое использовали, для того чтобы последний опорный элемент стал «малым». (Для более полного обсуждения этого примера см. Уилкинсон (1958а).)

Четвертый столбец табл. 10 дает вектор, полученный выполнением обратной подстановки с вектором e в качестве правой части. Можно видеть, что некоторые компоненты x очень большие; согласно нашим замечаниям предыдущего параграфа, это означает, что нормированный x и λ дают малый вектор-невязку. Следовательно, так как λ_1 хорошо отделено от λ_2 , нормированный x является хорошим собственным вектором. В действительности он верен почти с 7 десятичными знаками.

Заметим, что не все компоненты e были одинаково важны в обратной подстановке; последние 11 компонент совсем не существенны, и значения, которые они дают, становятся незначительными после нормировки. В действительности, если соответствующие вычисления выполняются по очереди для каждого собственного значения матрицы W_{21}^- , нахо-

дим, что для различных собственных векторов различные компоненты e играют основную роль. Характеристикой эффективности метода может служить то, что, работая с 26-разрядной двоичной арифметикой, каждый собственный вектор матрицы W_{21}^- был вычислен с 40 и более верными двоичными знаками после первой итерации и предельная точность была получена во второй итерации.

57. Естественно спросить, может ли процесс когда-нибудь оказаться несостоятельным и может ли быть получена последовательность значений, для которой (55.6) никогда не выполняется. Это кажется почти невозможным, так как может произойти лишь тогда, когда вектор b , который неявно используется в методе, содержит компоненту требуемого вектора порядка 2^{-t} или меньше, и ошибки округления таковы, что эта компонента не увеличивается.

Этот метод с интересной стороны освещает смысл термина «плохо обусловлена». Матрицы $(C - \lambda I)$, которые мы используем, почти вырожденные, так как λ является почти собственным значением. Поэтому численные решения, которые мы получаем из систем уравнений $(C - \lambda I)x_{r+1} = y_r$, в общем случае неверны даже в своих самых старших знаках. Тем не менее, в отношении нормированного вектора x_r процесс хорошо обусловлен, если только C не имеет более одного собственного значения, близкого к λ . Даже в этом случае ошибки в собственных векторах соответствуют влиянию малых возмущений в элементах C .

Близкие собственные значения и малые β_i

58. Из доказательств предыдущих параграфов мы видим, что собственный вектор предельной точности будет точным собственным вектором некоторой матрицы $(C + F)$, где F есть малая матрица возмущения такая, что

$$\|F\| < f(n) 2^{-t}, \quad (58.1)$$

причем $f(n)$ — некоторая медленно возрастающая функция от n , зависящая от типа используемой арифметики. Матрица F , вообще говоря, будет различной для каждого собственного значения. Предположим теперь, что C имеет два собственных значения λ_1 и λ_2 , разделенных величиной 2^{-p} , в то время как $\lambda_i - \lambda_1$ ($i \neq 1, 2$) не малы. Тогда мы можем предположить, что получим приближения v_1 и v_2 такие, что

$$v_1 = u_1 + f(n) 2^{-t} (2^p h_2 u_2 + \sum_{i=3}^n h_i u_i), \quad (58.2)$$

$$v_2 = u_2 + f(n) 2^{-t} (2^p k_1 u_1 + \sum_{i=3}^n k_i u_i), \quad (58.3)$$

где все h_i и k_i порядка единицы. Другими словами, v_1 и v_2 будут соответственно испорчены значительными компонентами по u_2 и u_1 , но незначительными компонентами по остальным u_i . Имеем

$$v_1^T v_2 = (k_1 + h_2) f(n) 2^{p-t} + 2^{-2t} (f(n))^2 \sum_{i=3}^n h_i k_i. \quad (58.4)$$

В общем случае h_2 и k_1 не будут связаны и, следовательно, когда p стремится к t и два собственных значения становятся близкими, v_1 и v_2 будут, вообще говоря, отличаться от ортогональных все больше и больше.

Мы можем сравнить это с тем положением, которое возникает при нахождении собственных векторов трехдиагональной матрицы методом Якоби. В этом случае мы должны бы иметь

$$v_1 = u_1 + 2^{-t} g(n) (2^p h_2 u_2 + \sum_{i=3}^n h_i u_i), \quad (58.5)$$

$$v_2 = u_2 + 2^{-t} g(n) (2^p k_1 u_1 + \sum_{i=3}^n k_i u_i), \quad (58.6)$$

где $g(n)$ — снова некоторая медленно возрастающая функция от n , несколько большая, чем $f(n)$, так как здесь несколько больше влияние ошибок округления. Однако из § 19 мы знаем, что v_1 и v_2 будут «почти ортогональны» и, следовательно, h_2 и k_1 почти равны по модулю и противоположны по знаку. Хотя при p , стремящемся к t , векторы v_1 и v_2 неизбежно становятся все более испорченными соответственно компонентами по u_2 и u_1 , это никоим образом не влияет на почти ортогональность.

Линейно независимые векторы, соответствующие совпадающим собственным значениям

59. Предположим теперь, что собственные значения так близки, что совпадают с рабочей точностью. Это имеет место, например, для λ_1 и λ_2 матрицы W_{21}^+ , если только мы не работаем с очень высокой точностью. Метод Якоби будет давать два почти ортогональных вектора, принадлежащих подпространству, натянутому на u_1 и u_2 ; эти векторы будут иметь очень малые компоненты по u_i ($i = 3, 4, \dots, n$), но они будут являться почти произвольными ортогональными комбинациями u_1 и u_2 .

С другой стороны, обратная итерация из §§ 53—55 позволяет получить лишь один собственный вектор для данного значения λ . Если случилось, что вычисленные λ_1 и λ_2 точно равны, то нам будет нужен некоторый метод вычисления второго вектора в подпространстве, натянутом на u_1 и u_2 . Теперь, когда C имеет патологически близкие собственные значения, будет обнаружено, что вектор, полученный по методу § 54, чрезвычайно чувствителен к тому значению λ , которое используется. Мы можем проиллюстрировать это с помощью простого примера.

Рассмотрим трехдиагональную матрицу

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \quad (59.1)$$

имеющую двукратное собственное значение $\lambda = 3$ с независимыми собственными векторами $(1, 0, 0)$ и $(0, 1, 1)$. Так как $\beta_2 = 0$, мы заменим его сначала на 10^{-t} , так что теперь все β_i не равны нулю. Если выполним один шаг обратной итерации с $\lambda = 3$, то получается вектор

$$\begin{bmatrix} 2 \times 10^t \\ 10^t + 1 \\ 10^t \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \quad (\text{после нормировки}). \quad (59.2)$$

С другой стороны, если возьмем $\lambda = 3 - 10^{-t}$, получается такой вектор:

$$\begin{bmatrix} 0 \\ 10^t \\ 10^t \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \text{ (после нормировки)}, \quad (59.3)$$

тогда как при $\lambda = 3 - 2 \times 10^{-t}$ имеем

$$\begin{bmatrix} -\left(\frac{1}{7}\right)10^t & -\frac{2}{7} \\ \left(\frac{2}{7}\right)10^t & +\frac{4}{7} \\ \left(\frac{2}{7}\right)10^t & \end{bmatrix} \Rightarrow \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 1 \end{bmatrix} \text{ (после нормировки)}. \quad (59.4)$$

Каждый из этих собственных векторов лежит в точном подпространстве, хотя очень малые изменения в λ приводят к совсем различным векторам. Третье собственное значение $\lambda = 1$ простое, и можно установить, что любое близкое к единице значение λ дает единственный собственный вектор $(0, 1, -1)$.

Большая чувствительность вычисленного собственного вектора к очень малым изменениям в значении λ может обернуться практическим преимуществом и может быть использована для получения независимых собственных векторов, соответствующих совпадающим или патологически близким собственным значениям. На АСЕ, работая с нормированными матрицами, мы искусственно разделяли все собственные значения по крайней мере на 3×2^{-t} . На практике это было весьма эффективно и всегда приводило к нужному числу независимых собственных векторов. Работая, например, с матрицей W_{21}^+ , мы получили таким образом полную систему независимых собственных векторов. Если мы имели собственное значение λ_i высокой кратности, то использовали λ_i , $\lambda_i \pm 3 \times 2^{-t}$, $\lambda_i \pm 6 \times 2^{-t}$, ... Возможно, стоит отметить, что использование, например, значения $\lambda_i \pm 9 \times 2^{-t}$ не имеет отрицательного влияния на точность вычисленного собственного вектора. Близость λ к собственному значению влияет на скорость сходимости, но не на достижимую точность обратной итерации. Конечно, мы не можем ожидать, что вычисленные векторы будут ортогональны, но вспомним, что ухудшение ортогональности обнаруживается уже тогда, когда собственные значения еще весьма далеки от патологически близких. Нужно признать, что это решение задачи не очень изящно. Оно имеет тот недостаток, что если, нам требуются ортогональные векторы, мы должны будем применить к вычисленным векторам процесс ортогонализации Шмидта. К тому же, если только процесс не может быть выполнен более обоснованно, остается опасность, что мы не получим полную цифровую информацию о подпространстве. Например, два вычисленных вектора предположительно могли быть такими:

$$v_1 = 0,6u_1 + 0,4u_2 + 2^{-t} \sum_{i=3}^n h_i u_i, \quad (59.5)$$

$$v_2 = 0,61u_1 + 0,39u_2 + 2^{-t} \sum_{i=3}^n k_i u_i, \quad (59.6)$$

где h_i и k_i порядка единицы. Математически они независимы, но направление, ортогональное первому вектору, определяется плохо. В действительности мы имеем

$$v_1 - v_2 = -0,01u_1 + 0,01u_2 + 2^{-t} \sum_{i=3}^n (h_i - k_i) u_i, \quad (59.7)$$

и, следовательно, в процессе ортогонализации будут потеряны два десятичных знака.

Иногда, когда исходная матрица имеет кратные собственные значения, матрица C , полученная из нее, будет иметь некоторые элементы β_i , почти равные нулю. Тогда мы можем расщепить C на прямую сумму ряда меньших трехдиагональных матриц и собственные векторы, соответствующие различным трехдиагональным матрицам, автоматически будут *точно* ортогональными. К сожалению, опыт свидетельствует, что мы не можем надеяться на появление малых β_i . Наоборот, матрица W_{21}^+ показывает, насколько далеко мы можем быть от той ситуации, в которой имеем явное разложение трехдиагональной матрицы. Тем не менее даже в этом случае, если мы разделим, например, матрицу на две матрицы 10 и 11 порядка, положив β_{11} равным нулю, то получим независимые ортогональные векторы, которые образуют подпространства, соответствующие патологически близким парам собственных значений, с очень высокой точностью. Поэтому вполне возможно, что такое разложение всегда допустимо, но тогда нужен некоторый надежный метод для решения вопроса, когда и как раскладывать.

Другой метод вычисления собственных векторов

60. По-видимому, в настоящее время метод обратной итерации, который мы описали, является самым хорошим из известных методов вычисления собственных векторов трехдиагональных матриц. Он в любом случае требует значительно меньшей работы, чем вычисление собственных значений. Далее, он дает ценную информацию о выполнении обратной итерации, что является очень важным для более общей проблемы собственных векторов.

Однако был предложен и ряд других методов, из которых следующий, принадлежащий Деккеру (1962), возможно, самый удачный. Он основан по существу на процессе, описанном в § 50. Берем $x_1 = 1$ и последовательно вычисляем $x_2, x_3, \dots$, используя уравнения (48.1) и (48.2). Этот процесс продолжается, пока выполняется условие $|x_{r+1}| \geq |x_r|$. Если мы достигнем этапа, на котором это условие не выполнено, то берем $x'_n = 1$ и последовательно вычисляем $x'_{n-1}, x'_{n-2}, \dots$ из уравнений (48.3) и (48.2). Опять продолжаем до тех пор, пока не выполнится неравенство $|x'_{s-1}| < |x'_s|$, затем возвращаемся к первому процессу и снова продолжаем до тех пор, пока не произойдет уменьшение.

Если мы имеем $|x_1| > |x_2| > |x_3| > \dots$ и $|x'_n| > |x'_{n-1}| > |x'_{n-2}| > \dots$, то будем переключаться попеременно с одного конца на другой. Если же мы, например, имеем

$$|x_1| > |x_2| > |x_3| \leq |x_4| \leq |x_5| \leq \dots \leq |x_n|, \quad (60.1)$$

и

$$|x'_n| > |x'_{n-1}| > |x'_{n-2}|, \quad (60.2)$$

то будем переключаться назад и вперед дважды, но после возвращения к первой подстановке во второй раз $|x_i|$ увеличиваются, и мы будем продолжать процесс с первой подстановкой все остальное время, отбрасывая значения x'_i . Аналогично мы в конце концов можем отбросить значения x_i . Мы можем, напротив, встретиться во внутренней точке, когда при уменьшающихся последовательностях x_i , и x'_i впервые произойдет наложение. В этом случае две последовательности объединяются, как описано в § 50.

По-видимому этот процесс должен давать хорошие результаты, но, похоже, что он будет не очень точным, когда прямая и обратная последовательности быстро уменьшаются в том месте, где они пересекаются. К тому же кажется, что процесс нелегко изменить так, чтобы получить независимые векторы, когда C имеет патологически близкие собственные значения, но не имеет значений β_i , которые дают ясное указание на то, как расщепить матрицу. Ряд вариантов прямой и обратной подстановок был исследован автором и другими.

Численный пример

61. В качестве иллюстрации поведения варианта Деккера рассмотрим вычисление собственного вектора W_{21}^+ , соответствующего собственному значению λ_1 , которое патологически близко к λ_2 . Эта матрица симметрична относительно второй диагонали, поэтому прямая и обратная последовательности совпадают. Обе последовательности вначале уменьшаются, поэтому мы переключаемся попеременно то на прямую, то на обратную последовательность. Когда мы доходим до x_9 , прямая последовательность начинает расти и продолжает возрастать как раз до x_{21} . Поэтому обратная последовательность отбрасывается. Эти результаты приводятся в табл. 11. На первый взгляд они кажутся несколько расхолаживающими,

Таблица 11

$$\lambda = 10,74619\,419$$

$x_1 = 1,00000\,000$	$x_{12} = 0,17133\,492$	$x'_{21} = 1,00000\,000$
$x_2 = 0,74619\,419$	$x_{13} = 1,65376\,40$	$x'_{20} = 0,74619\,419$
$x_3 = 0,30299\,996$	$x_{14} = 1,42928\,06 \times 10$	$x'_{19} = 0,30299\,996$
$x_4 = 0,08590\,254$	$x_{15} = 1,09061\,09 \times 10^2$	$x'_{18} = 0,08590\,254$
$x_5 = 0,01880\,764$	$x_{16} = 7,21454\,49 \times 10^2$	$x'_{17} = 0,01880\,764$
$x_6 = 0,00336\,217$	$x_{17} = 4,03655\,65 \times 10^3$	$x'_{16} = 0,00336\,217$
$x_7 = 0,00051\,204$	$x_{18} = 1,84368\,27 \times 10^4$	$x'_{15} = 0,00051\,204$
$x_8 = 0,00009\,215$	$x_{19} = 6,50313\,78 \times 10^4$	$x'_{14} = 0,00009\,215$
$x_9 = 0,00020\,177$	$x_{20} = 1,60151\,97 \times 10^5$	
$x_{10} = 0,00167\,257$	$x_{21} = 2,14625\,06 \times 10^5$	
$x_{11} = 0,01609\,942$		

Нормированный вектор

$x_1 = 0,00000\,466$	$x_8 = 0,00000\,000$	$x_{15} = 0,00050\,815$
$x_2 = 0,00000\,348$	$x_9 = 0,00000\,000$	$x_{16} = 0,00336\,146$
$x_3 = 0,00000\,141$	$x_{10} = 0,00000\,001$	$x_{17} = 0,01880\,748$
$x_4 = 0,00000\,040$	$x_{11} = 0,00000\,007$	$x_{18} = 0,08590\,249$
$x_5 = 0,00000\,009$	$x_{12} = 0,00000\,080$	$x_{19} = 0,30299\,993$
$x_6 = 0,00000\,002$	$x_{13} = 0,00000\,771$	$x_{20} = 0,74619\,418$
$x_7 = 0,00000\,000$	$x_{14} = 0,00006\,659$	$x_{21} = 1,00000\,000$

так как действительное значение x_{21} , очевидно, полностью определяется ошибкой в λ и ошибками округления. Однако x_{20} и x_{21} удовлетворяют двадцать первому уравнению почти точно и, следовательно, нормированный вектор, полученный из x , должен дать вектор-невязку с компонентами

порядка 10^{-8} и поэтому в пределах рабочей точности он лежит в подпространстве, натянутом на первые два собственных вектора. Мы даем нормированный собственный вектор. Если u_1 и u_2 являются первыми двумя нормированными (по ∞ -норме) собственными векторами матрицы W_{21}^+ , то мы проверили, что

$$x = 0,50000\ 233u_1 + 0,49999\ 767u_2 + O(10^{-8}). \quad (61.1)$$

Если мы попытаемся вычислить другой вектор, беря несколько иное значение λ , то найдем, что снова принимаем прямую последовательность, но x_{21} будет полностью отличаться от своего прежнего значения. К сожалению, этот новый вектор $\bar{x}$ таков, что

$$\bar{x} = (0,5 + \varepsilon)u_1 + (0,5 - \varepsilon)u_2 + O(10^{-8}), \quad (61.2)$$

где ε снова порядка 10^{-5} . Поэтому мы не получили точной информации об ортогональном к x направлении в подпространстве, натянутом на u_1 и u_2 . Однако интересно отметить, что если, используя первоначальное значение λ , мы начнем с обратной последовательности вместо прямой, то в конце концов перейдем к обратной последовательности, а она совпадает с прямой последовательностью. Поэтому вычисленный вектор таков:

$$0,50000\ 233u_1 - 0,49999\ 767u_2 + O(10^{-8}), \quad (61.3)$$

и он почти ортогонален вектору x из (61.1). Таким приемом мы получаем полную информацию о подпространстве из одного значения λ .

Замечания о проблеме собственных значений для трехдиагональных матриц

62. Мы обсудили проблему собственных значений для трехдиагональных матриц довольно подробно потому, что она не только важна сама по себе, но дает также поучительные примеры различий между классическим и численным анализом. Предельные процессы в тех методах, которые мы обсудили, в некоторых отношениях являются плохим руководством для определения того, что можно ожидать на практике, когда «малость» величин строго ограничена точностью вычислений. Приведенные примеры не следует рассматривать как простую численную иллюстрацию; читателю рекомендуется внимательно их изучить и иметь в виду, если он попытается разработать метод вычисления собственных векторов более эффективный, чем тот, который дали мы.

Завершение методов Гивенса и Хаусхолдера

63. Для того чтобы завершить методы Гивенса и Хаусхолдера, мы должны получить собственные векторы исходной матрицы из собственных векторов трехдиагональной матрицы. Для метода Гивенса это почти идентично вычислению, описанному в § 16 в связи с методом Якоби. Если x есть собственный вектор матрицы C , то z , определенный равенством

$$z = R_1^T R_2^T \dots R_s^T x, \quad (63.4)$$

где R_i являются матрицами преобразования, есть собственный вектор матрицы A_0 . В общем случае имеется $\frac{1}{2}(n-1)(n-2)$ вращений и вычи-

сленние каждого вектора z требует $2(n-1)(n-2)$ умножений. Если нужны все векторы, то эта часть вычислений требует около $2n^3$ умножений.

Для метода Хаусхолдера

$$z = P_1 P_2 \dots P_{n-2} x, \quad (63.2)$$

где P_r суть матрицы преобразования. Естественно, что каждая P_r используется индивидуально в своей факторизованной форме $I - u_r u_r^T / 2K_r^2$ или $I - 2y_r y_r^T / \|y_r\|^2$ согласно формулам § 30 или § 33.

Если мы обозначим $x = x^{(n-2)}$ и $x^{(r-1)} = P_r x^{(r)}$, то для первой формулы имеем

$$x^{(r-1)} = x^{(r)} - \left(\frac{u_r^T x^{(r)}}{2K_r^2} \right) u_r, \quad z = x^{(0)}. \quad (63.3)$$

Для восстановления каждого вектора требуется около n^2 умножений и, следовательно, n^3 умножений, если нужны все n векторов. Заметим, что как для метода Гивенса, так и для метода Хаусхолдера восстановление всех векторов требует больше вычислений, чем приведение матрицы к трехдиагональной форме.

64. Восстановление начальных векторов численно устойчиво и не представляет какой-либо особенно интересной задачи; почти все, что нам требуется, охвачено анализом, который мы дали в главе 3. Мы можем излагать методы Гивенса и Хаусхолдера одновременно.

Пусть R означает точное произведение точных ортогональных матриц, соответствующих вычисленным A_r . Тогда имеем

$$C \equiv R A_0 R^T + F, \quad (64.1)$$

где для различных методов была найдена оценка F . Если x есть вычисленный нормированный собственный вектор матрицы C , то

$$Cx - \lambda x = \eta, \quad (64.2)$$

где анализ ошибок дает оценку для $\|\eta\|_2$. Следовательно,

$$(C + G)x = \lambda x, \quad (64.3)$$

где G может быть взята симметричной, если λ есть отношение Релея, соответствующее x , и из гл. 3, § 56 имеем

$$\|G\|_2 = \|\eta\|_2. \quad (64.4)$$

Наконец, анализ гл. 3, §§ 23, 34 и 45 дает оценку вектора f , определенного через вычисленный вектор z равенством

$$z \equiv R^T x + f. \quad (64.5)$$

Объединяя (64.1) и (64.3), получаем

$$(A_0 + R^T F R + R^T G R) R^T x = \lambda R^T x. \quad (64.6)$$

Итак, $R^T x$ есть точный собственный вектор симметричной матрицы, отличающейся от A_0 на возмущение, ограниченное по норме суммой $\|F\| + \|G\|$, в то время как z отличается от $R^T x$ на f . Оценка для f не зависит от обусловленности задачи определения собственных векторов.

Главная опасность для плохо обусловленных векторов исходит из возможного влияния симметричного возмущения A_0 . Заметим, что G зависит только от того метода, который мы используем для нахождения собствен-

ных векторов матрицы C , тогда как F и f зависят от метода трехдиагонализации. За исключением, возможно, метода Хаусхолдера, выполненного в плавающей запятой с накоплением, оценка ошибки вектора определяется оценкой для F . Для метода Хаусхолдера оценка F настолько хороша, что ошибка за счет G становится сравнимой.

Сравнение методов

65. Если нас интересуют лишь собственные значения, то метод Якоби практически не может сравниться ни с методом Гивенса, ни с методом Хаусхолдера. Предполагая, что было сделано 6 циклов, мы заключаем, что приведение к диагональной форме по методу Якоби требует в 9 раз больше времени, чем трехдиагонализация Гивенса, и в 18 раз больше, чем трехдиагонализация Хаусхолдера. Если нужны лишь некоторые собственные значения, то метод Якоби сравнивать особенно невыгодно. Так как преобразование Хаусхолдера вдвое быстрее преобразования Гивенса и для информации относительно преобразования нужен вдвое меньший объем памяти, то метод Хаусхолдера, по-видимому, должен быть самым эффективным.

Все три метода настолько точны на практике, что точность едва ли должна быть решающим фактором, но в любом случае соотношение в пользу методов Хаусхолдера и Гивенса. Если могут быть накоплены скалярные произведения, то метод Хаусхолдера чрезвычайно точен.

Есть одно небольшое замечание в пользу метода Гивенса. Если элемент, который должен быть исключен, уже равен нулю, то соответствующее вращение легко опускается; верно, что в этом случае один элемент из w_i равен нулю и в методе Хаусхолдера, но преимущество этого не так легко использовать. Если исходная матрица содержит много нулевых элементов, то число необходимых вращений значительно уменьшается и, хотя некоторые нули постепенно заменяются ненулевыми элементами по мере продолжения процесса, более поздние вращения требуют меньше вычислений, чем более ранние.

Если нужны собственные векторы, то в пользу метода Якоби должно быть сказано несколько больше. Хотя собственный вектор, вычисленный по методу Якоби, будет иметь в общем случае большую компоненту, ортогональную к соответствующему подпространству, чем аналогичный вектор, вычисленный по методу Гивенса или Хаусхолдера и методу §§ 53—55, все же метод Якоби имеет то преимущество, что он дает почти ортогональные векторы. Чрезвычайная простота метода Якоби с точки зрения программирования может также быть причиной его использования. Однако в целом я чувствовал, что, даже когда требуются собственные векторы, метод Хаусхолдера кажется более предпочтительным; недостатки метода определения независимых векторов, соответствующих близким собственным значениям, не такие серьезные, чтобы оправдывать использование более медленного и менее точного метода.

Квазисимметричные трехдиагональные матрицы

66. Элементы общей трехдиагональной матрицы C могут быть определены равенствами

$$c_{ii} = \alpha_i, \quad c_{i, i+1} = \beta_{i+1}, \quad c_{i+1, i} = \gamma_{i+1}, \quad c_{ij} = 0 \text{ в остальных случаях.} \quad (66.1)$$

В общем случае собственные значения такой матрицы могут быть комплексными, даже если элементы вещественные. Однако если α_i

вещественные и

$$\beta_i \gamma_i > 0 \quad (i = 2, \dots, n), \quad (66.2)$$

то C может быть преобразована в вещественную симметричную матрицу посредством подобного преобразования с диагональной матрицей D . Если мы определим D соотношениями

$$d_{1,1} = 1, \quad d_{ii} = (\gamma_2 \gamma_3 \dots \gamma_i / \beta_2 \beta_3 \dots \beta_i)^{1/2}, \quad (66.3)$$

то

$$D^{-1}CD = T, \quad (66.4)$$

где T — трехдиагональная матрица и

$$t_{ii} = \alpha_i, \quad t_{i,i+1} = t_{i+1,i} = (\beta_{i+1} \gamma_{i+1})^{1/2}. \quad (66.5)$$

Если β_i или γ_i равны нулю, то собственные значения являются собственными значениями ряда меньших трехдиагональных матриц, так что этот случай не представляет каких-либо трудностей.

Мы можем использовать свойство последовательности Штурма для того, чтобы вычислить собственные значения так же, как в §§ 37—39; нужно заменить лишь β_i^2 на $\beta_i \gamma_i$. Трехдиагональные матрицы, имеющие вещественные α_i и с элементами β_i и γ_i , удовлетворяющими соотношению (66.2), иногда называются *квазисимметричными* трехдиагональными матрицами.

Вычисление собственных векторов

67. Метод вычисления собственных векторов, описанный в §§ 53—55, не учитывает симметрию, и в действительности симметрия, как правило, нарушается при исключении Гаусса с перестановками. Поэтому метод может быть сразу же применен к квазисимметричным матрицам. В последующих главах этой книги мы будем иметь дело с собственными векторами трехдиагональных матриц, которые получаются с помощью подобных преобразований из более общих матриц. Таким образом, будем иметь

$$X^{-1}AX = C \quad (67.1)$$

для некоторой невырожденной матрицы X . Соответственно собственному вектору y матрицы C нас будет интересовать вектор x , определенный соотношением

$$x = Xy, \quad (67.2)$$

так как он будет собственным вектором матрицы A . Если элементы β_i и γ_i имеют большой разброс порядков величин, то это обычно происходит из-за матрицы X , имеющей очень различные по величине столбцы. Если X заменить на Y , определенную равенством

$$Y = XD, \quad (67.3)$$

где D есть любая невырожденная диагональная матрица, то

$$Y^{-1}AY = D^{-1}CD = T, \quad (67.4)$$

где T — также трехдиагональная матрица. Изменения в матрице D приводят к изменениям в перестановках, которые используются в исключении Гаусса для матрицы $(T - \lambda I)$, а также повлияют на эффективность

вектора e , который мы использовали в качестве правой части для обратной подстановки в § 54. Вообще говоря, кажется целесообразным выбирать D так, чтобы столбцы Y имели один и тот же порядок величин, и тогда применять к полученной матрице T метод §§ 53—55.

Уравнения вида $Ax = \lambda Bx$ и $ABx = \lambda x$

68. В главе 1, §§ 31—33 мы рассматривали уравнения вида

$$Ax = \lambda Bx \quad (68.1)$$

и

$$ABx = \lambda x, \quad (68.2)$$

где A и B — вещественные и симметричные и по крайней мере одна из них положительно определенная.

Мы показали, что если, например, B — положительно определенная и мы решили ее полную проблему собственных значений, так что

$$R^T B R = \text{diag}(\beta_i^2) = D^2 \quad (68.3)$$

для некоторой ортогональной R , то (68.1) эквивалентно решению обычного уравнения

$$Pz = \lambda z, \quad (68.4)$$

где

$$P = D^{-1} R^T A R D^{-1} \quad \text{и} \quad z = D R^T x. \quad (68.5)$$

Если мы решали полную проблему для матрицы B методом Якоби, Гивенса или Хаусхолдера, то будем иметь R в факторизованной форме, по которой может быть вычислена вещественная симметричная матрица P . Такие методы в действительности использовались для решения (68.1) и соответствующие методы для решения (68.2). Они численно устойчивы и имеют то преимущество, что для решения полной проблемы матриц B и P дважды используется одна и та же программа. Однако объем работы значителен.

69. Следующий метод численно даже более устойчив и требует существенно меньшей работы. Для сравнения рассмотрим уравнение (68.2). Если B — положительно определенная, то с помощью разложения Холецкого имеем

$$B = LL^T \quad (69.1)$$

с вещественной невырожденной нижней треугольной матрицей L . Следовательно, (68.2) может быть записано в виде

$$ALL^T x = \lambda x, \quad (69.2)$$

откуда

$$(L^T A L) L^T x = \lambda L^T x. \quad (69.3)$$

Поэтому, если λ и y суть собственное значение и собственный вектор матрицы $L^T A L$, то λ и $(L^T)^{-1}y$ являются решениями уравнения (68.2).

Теперь мы знаем, что разложение Холецкого для матрицы B в высшей степени устойчиво, особенно если мы можем накапливать скалярные произведения. Далее, оно требует только $n^3/6$ умножений. При вычисле-

нии $L^T A L$ по L мы можем полностью использовать симметрию. Если обозначим

$$F = A L, \quad (69.4)$$

$$G = L^T F, \quad (69.5)$$

то G есть симметричная матрица $L^T A L$, и, следовательно, нам нужен только нижний треугольник ее элементов. Но L^T — верхняя треугольная, поэтому при вычислении нижнего треугольника G используется лишь нижний треугольник F . Следовательно, нам нужно вычислить только нижний треугольник матрицы F из (69.4). Для его определения требуется $\frac{1}{3} n^3$ умножений и еще $\frac{1}{6} n^3$ умножений для определения нижнего

треугольника матрицы G . Итак, G получается из A и B за $\frac{2}{3} n^3$ умножений! Если на всех соответствующих этапах накапливаются скалярные произведения, то общее влияние ошибок округления будет удерживаться на очень низком уровне.

Обращаясь к уравнению (68.1), находим, что теперь требуется вычисление матрицы $(L^{-1}) A (L^{-1})^T$ и решение треугольных систем уравнений. Мы снова можем накапливать скалярные произведения на всех соответствующих этапах и использовать симметрию окончательной матрицы. Здесь мы извлекаем существенную выгоду из высокой точности решения треугольных систем уравнений.

Когда B плохо обусловлена в отношении обращения, то этот процесс для $(A - \lambda B)$ будет сам плохо обусловленным. Если мы предположим (без существенной потери общности), что $\|B\|_2 = 1$, то в плохо обусловленном случае $\|L^{-1} A (L^{-1})^T\|_2$ будет намного больше чем $\|A\|_2$. Вообще говоря, исходное уравнение будет иметь несколько «больших» собственных значений, а остальные «нормальной величины»; последние будут определяться неточно *по этому процессу*. В пределе, когда B имеет нулевое собственное значение кратности r , определитель матрицы $(A - \lambda B)$ будет полиномом степени $(n - r)$; поэтому $(A - \lambda B)$ имеет r «бесконечных» собственных значений. Тем не менее конечные собственные значения матрицы $(A - \lambda B)$ в общем случае не будут чувствительны к возмущениям в A и B , и их плохое определение должно рассматриваться как слабость метода. Аналогичная трудность возникает и в том случае, когда B диагонализируется, так как тогда D имеет r нулевых элементов. (Дальнейшие замечания см. в конце этой главы.)

Численный пример

70. В табл. 12 мы показываем вычисление как $L^T A L$, так и $L^{-1} A (L^{-1})^T$, соответствующее одной и той же паре матриц A и B . Всюду использовалась 6-разрядная десятичная арифметика с накоплением скалярных произведений. Для того чтобы найти нули определителей матриц $(AB - \lambda I)$ и $(A - \lambda B)$, нужны собственные значения симметричных матриц $L^T A L$ и $L^{-1} A (L^{-1})^T$. Почти всюду мы дали только один треугольник симметричных матриц; исключение составляет матрица $W = L^{-1} A (L^{-1})^T$. Эта матрица получается решением уравнения $LW = Z$ и при вычислении r -го столбца W мы используем предварительно вычисленные значения $W_{r1}, W_{r2}, \dots, W_{r, r-1}$ в качестве значений $W_{1r}, W_{2r}, \dots, W_{r-1, r}$. Для проведения процесса в обоих случаях требуется выполнить очень мало работы.

Таблица 12
В (симметричная и положительно определенная)

$$\begin{bmatrix} 0,983 & 0,105 & 0,213 & 0,122 \\ & 0,807 & 0,214 & 0,132 \\ & & 0,903 & 0,213 \\ & & & 0,977 \end{bmatrix}$$

L^T (где $B = LL^T$)

$$\begin{bmatrix} 0,991464 & 0,466421 & 0,214834 & 0,123050 \\ & 0,932365 & 0,491177 & 0,119612 \\ & & 0,905703 & 0,180741 \\ & & & 0,956496 \end{bmatrix}$$

$G = L^T F = L^T A L$ (симметричная)

$$\begin{bmatrix} 1,408298 & & & \\ 0,854808 & 0,416283 & & \\ 0,499655 & 0,480942 & 0,576304 & \\ 0,570771 & 0,404733 & 0,436495 & 0,288188 \end{bmatrix}$$

$W = L^{-1} Z = L^{-1} A (L^{-1})^T$ (симметричная)

$$\begin{bmatrix} 0,954169 & 0,493351 & -0,088100 & 0,268090 \\ 0,493351 & 0,042050 & 0,206361 & 0,176732 \\ -0,088100 & 0,206361 & 0,476486 & 0,241206 \\ 0,268090 & 0,176732 & 0,241206 & 0,085213 \end{bmatrix}$$

$F = A L$ (только нижний треугольник)

$$\begin{bmatrix} 1,126474 & & & \\ 0,751562 & 0,300629 & & \\ 0,432593 & 0,446574 & 0,545238 & \\ 0,596731 & 0,423141 & 0,456348 & 0,301296 \end{bmatrix}$$

$Z^T = L^{-1} A$ (только верхний треугольник)

$$\begin{bmatrix} 0,943050 & 0,618278 & 0,218868 & 0,416556 \\ & 0,121310 & 0,300929 & 0,272078 \\ & & 0,452079 & 0,330676 \\ & & & 0,179229 \end{bmatrix}$$

Для сравнения матрицы $L^T A L$ и $L^{-1} A (L^{-1})^T$ были вычислены, используя много большую точность. Максимальная ошибка в элементах представленной матрицы $L^T A L$ равна 10^{-6} (2,04), и она имеет место только в одном элементе, да и то в результате исключительно неудачных округлений; все остальные элементы имеют ошибки меньше, чем 10^{-6} . Максимальная ошибка в элементах вычисленной матрицы $L^{-1} A (L^{-1})^T$ равна 10^{-6} (0,98). Эти результаты подтверждают замечательную численную устойчивость обоих процессов.

Одновременное приведение A и B к диагональному виду

71. Из решения полной проблемы для уравнения

$$Ax = \lambda Bx \quad (71.1)$$

можем получить такую невырожденную матрицу X , что

$$X^T A X = \text{diag}(\lambda_i), \quad X^T B X = I. \quad (71.2)$$

Для этого рассмотрим эквивалентное уравнение

$$L^{-1} A (L^{-1})^T y = \lambda y \quad (71.3)$$

с симметричной матрицей, и пусть $y_1, y_2, \dots, y_n$ есть ортонормальная система собственных векторов. Тогда имеем

$$Y^T Y = I \quad (71.4)$$

и

$$L^{-1} A (L^{-1})^T Y = Y \text{diag}(\lambda_i). \quad (71.5)$$

Если мы определим X соотношением

$$Y = L^T X, \quad (71.6)$$

то X есть невырожденная матрица и равенство (71.4) становится таким:

$$X^T L L^T X = I, \quad \text{т. е.} \quad X^T B X = I. \quad (71.7)$$

С другой стороны, умножая (71.5) слева на Y^T , имеем

$$X^T A X = \text{diag}(\lambda_i). \quad (71.8)$$

Трехдиагональные A и B

72. Если A и B — трехдиагональные матрицы, то метод, который мы обсудили для решения уравнения $Ax = \lambda Bx$, становится непривлекательным, так как $L^{-1} A (L^{-1})^T$, вообще говоря, будет полной матрицей. Эта задача может быть решена методом, аналогичным тому, который использовался в § 39. Сначала докажем следующую простую лемму.

Если A и B симметричные и B положительно определенная, то нули определителя матрицы $(A_r - \lambda B_r)$ разделяют нули определителя матрицы $(A_{r+1} - \lambda B_{r+1})$, где A_i и B_i — матрицы главных миноров порядка i . Это следует из замечания, что если

$$y = L^T x, \quad (72.1)$$

то

$$\frac{y^T L^{-1} A (L^{-1})^T y}{y^T y} = \frac{x^T A x}{x^T L L^T x} = \frac{x^T A x}{x^T B x}, \quad (72.2)$$

и уравнение (72.1) дает взаимно однозначное соответствие между ненулевыми x и y . Следовательно, нули определителя матрицы $(A - \lambda B)$ даются экстремумами отношения $x^T A x / x^T B x$ для ненулевых x (ср. гл. 2, § 43). Теперь мы можем доказать теорему разделения для нулей определителей матриц $(A_r - \lambda B_r)$ точно таким же путем, как в гл. 2, § 47. Итак, если обозначим

$$a_{ii} = \alpha_i, \quad a_{i+1, i} = \alpha_i, \quad b_{ii} = \alpha'_i, \quad b_{i+1, i} = \alpha'_{i+1}, \quad (72.3)$$

то полиномы, определенные соотношениями

$$\left. \begin{aligned} p_0(\lambda) &= 1, \quad p_1(\lambda) = \alpha_1 - \lambda \alpha'_1, \\ p_r(\lambda) &= (\alpha_r - \lambda \alpha'_r) p_{r-1}(\lambda) - (\beta_r - \lambda \beta'_r)^2 p_{r-2}(\lambda), \end{aligned} \right\} \quad (72.4)$$

образуют последовательность Штурма, и мы можем найти нули $p_n(\lambda)$ методом деления пополам так же, как в § 39. Мы можем найти знаки определителей матриц $(A_r - \lambda B_r)$ ($r = 1, \dots, n$) так же, как в § 47, выполняя исключение Гаусса с перестановками в трехдиагональной матрице $(A - \lambda B)$. Как и в более простой задаче, нам не обязательно все время следовать методу деления пополам, а можно переключаться на некоторый квадратично сходящийся процесс для нахождения нулей $p_r(\lambda) = \det(A - \lambda B)$, если только они изолированы.

Для того чтобы вычислить собственный вектор при хорошем приближении λ , обратная итерация теперь требует решения уравнений

$$(A - \lambda B) x_{r+1} = y_r, \quad (72.5)$$

$$y_{r+1} = B x_{r+1}, \quad (72.6)$$

так как они означают, что

$$[L^{-1}A(L^{-1})^T - \lambda I] L^T x_{r+1} = L^T x_r. \quad (72.7)$$

Снова в общем случае мы будем ожидать, что достаточно двух итераций.

Эти методы полностью используют трехдиагональный вид A и B .

Комплексные эрмитовы матрицы

73. Методы, которые мы описали для вещественных симметричных матриц, имеют точные аналогии для комплексных эрмитовых матриц. Для методов Якоби и Гивенса, с одной стороны, и Хаусхолдера, с другой, нам потребуются комплексные унитарные матрицы из гл. 1, §§ 43, 44 и §§ 45, 46 соответственно. Наиболее естественное расширение формул, использованных в § 29, дает метод, в котором для определения каждой элементарной эрмитовой матрицы нужно два извлечения квадратного корня. Небольшая модификация уменьшает это требование до одного извлечения квадратного корня. В обозначениях § 29 мы теперь имеем

$$\left. \begin{aligned} P_r &= I - u_r u_r^H / 2K_r^2, & u_{ir} &= 0 \quad (i = 1, 2, \dots, r), \\ u_{r+1, r} &= a_{r, r+1} (1 + S_r^2 / T_r), & u_{ir} &= a_{ri} \quad (i = r+2, \dots, n), \\ S_r^2 &= \sum_{i=r+1}^n |a_{ri}|^2, & T_r &= (|a_{r, r+1}|^2 S_r^2)^{1/2}, \\ 2K_r^2 &= S_r^2 + T_r, \end{aligned} \right\} \quad (73.1)$$

где мы выбрали знаки из соображений устойчивости.

Матрицы остаются эрмитовыми на всех этапах, какой бы из трех методов мы ни использовали. Метод Якоби дает в качестве предельной вещественную диагональную матрицу; согласно гл. 1, § 44 может быть показано, что и метод Гивенса дает вещественную трехдиагональную матрицу. Метод Хаусхолдера дает трехдиагональную эрмитову матрицу C , которая, вообще говоря, будет иметь комплексные внедиагональные элементы. Это следует из замечания в начале § 46 гл. 1. Мы имеем

$$\left. \begin{aligned} c_{ii} &= \alpha_i \quad (\text{вещественное}), \\ c_{i, i+1} &= \beta_{i+1}, \quad c_{i+1, i} = \bar{\beta}_{i+1} \end{aligned} \right\} \quad (73.2)$$

и, следовательно,

$$c_{i, i+1}c_{i+1, i} = |\beta_{i+1}|^2 \quad (\text{вещественное}). \quad (73.3)$$

Поэтому то, что β_i комплексные, не имеет большого значения.

Оказалось неожиданным то, что мы встречали мало примеров комплексных эрмитовых матриц, несмотря на их важность в теоретической физике, и комплексные варианты наших программ не были сделаны. Как отмечалось в гл. 3, § 56, полная проблема для комплексной эрмитовой матрицы $(B + iC)$ может быть сведена к полной проблеме для вещественной симметричной матрицы A , где

$$A = \begin{bmatrix} B & -C \\ C & B \end{bmatrix}, \quad (73.4)$$

и на ACE и DEUCE мы это делали. Мы знаем, что собственные значения A суть соответственно $\lambda_1, \lambda_1, \lambda_2, \lambda_2, \dots, \lambda_n, \lambda_n$, так что нам нужно лишь найти 1-е, 3-е, $\dots$, $(2n - 1)$ -е собственные значения трехдиагональной матрицы, полученной из A . Наличие совпадающих собственных значений не влияет на точность, с которой они определяются, и нужен лишь один вектор в подпространстве собственных векторов, соответствующих каждому двойному собственному значению. Все векторы в этом подпространстве соответствуют одному и тому же комплексному собственному вектору матрицы $B + iC$. Трехдиагональные матрицы, полученные по методам Гивенса или Хаусхолдера, должны расщепляться на две равные матрицы, но на практике это может происходить не всегда.

Однако матрица A требует вдвое больше памяти, чем $(B + iC)$, и так как число умножений в преобразованиях как Гивенса, так и Хаусхолдера изменяется как куб порядка, то в преобразовании A будет в восемь раз больше вещественных умножений, чем комплексных умножений в преобразовании $(B + iC)$. Так как одно комплексное умножение включает четыре вещественных умножения, то преобразование $(B + iC)$ содержит лишь половину вычислений преобразования A по тому же самому методу. В преобразовании Гивенса два из четырех элементов каждого плоского вращения вещественные, за исключением элементов в плоскостях $(i, i + 1)$, и если мы используем это, то число умножений в преобразовании матрицы $(B + iC)$ уменьшится в $3/4$ раза. Если требовалось большое число таких вычислений и если рассматриваемые матрицы были так велики, что становилась важной проблема размещения их в памяти, то эта причина могла бы стать очень веской для того, чтобы строить более сложные комплексные варианты.

Дополнительные замечания

Настоящий интерес к методу Якоби появляется в основном после его переоткрытия в 1949 г. Голдштейном, Мюрреем и Нейманом, хотя он использовался на настоящих вычислительных машинах в Национальной физической лаборатории в 1947 г. Необходимость записывать много результатов промежуточных вычислений делает метод Якоби существенно более эффективным на быстродействующих вычислительных машинах. С 1949 г. метод получил широкое освещение в литературе; мы упомянем статьи Голдштейна, Мюррея и Неймана (1959), Грегори (1953), Хенричи (1958а), Попе и Томпкинса (1957), Шенхаге (1961) и Уилкинсона (1962с). Статья Голдштейна и др. дает анализ ошибок процесса; большая часть работы была выполнена к 1951 г., хотя не публиковалась до 1959 г.

Общее превосходство метода Гивенса над методом Якоби для определения собственных значений получало признание сравнительно медленно, несмотря на то, что Гивенс (1954) дал строгий *априорный* анализ ошибок, который, по моему мнению, является одним из поворотных пунктов в истории этого вопроса. Вероятно, факт, что методы, предложенные для вычисления собственных векторов, были найдены ненадежными, привел к неоправданному подозрению в отношении точности собственных значений.

Впервые Хаусхолдер предложил использование матриц отражения для приведения к трехдиагональной форме в лекции, прочитанной в Урбане в 1958 г., и снова коротко упоминает об этом в совместной статье с Бауэром (1959). Общее превосходство этого метода над методом Гивенса как по скорости, так и по точности было впервые обнаружено Уилкинсоном (1960а). Анализ ошибок для преобразований с фиксированной запятой был дан Уилкинсоном (1962b), для преобразований с плавающей запятой — Ортега (1963).

Вычисление собственных значений трехдиагональных матриц, использующее свойство последовательности Штурма, и анализ ошибок были описаны Гивенсом (*указанная работа*). Этот анализ, относящийся к вычислению с фиксированной запятой и специальным нормированием, и был первым анализом, в котором явно сказано о том, что теперь известно как обратный анализ ошибок, хотя неявно эта идея присутствовала в статьях Неймана и Голдштейна (1947) и Тюринга (1948).

Задача вычисления собственных векторов трехдиагональной матрицы, когда известны хорошие приближения собственных значений, была рассмотрена Брукером и Сумнером (1956), Форсайтом (1958), Уилкинсоном (1958а) и Ланцосом в статье Россера и др. (1951). Обратная итерация дает очень удовлетворительное решение задачи, когда речь идет о достаточно хорошо разделенных собственных значениях.

Задача определения полной надежной цифровой информации в подпространстве, натянутом на собственные векторы, соответствующие совпадающим или патологически близким собственным значениям, никогда не была хорошо решена. Аналогичная задача имеет значение также в связи с определением собственных векторов матриц Хессенберга и обсуждается в главе 9.

Решение полной проблемы для $(A - \lambda B)$ недостаточно удовлетворительно, когда B плохо обусловлена в отношении обращения. То, что конечные собственные значения могут быть хорошо определены даже тогда, когда B вырождена, иллюстрируется на примере матриц

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Мы имеем $\det(A - \lambda B) = 3 - 2\lambda$, так что собственные значения суть $\frac{1}{2}$ и ∞ . Малые изменения в A и B вносят малые изменения в конечное собственное значение, но бесконечное становится конечным, хотя и очень большим.

Если A хорошо обусловлена и положительно или отрицательно определена, то мы можем соответственно работать с $(B - \lambda A)$ или $(-B - \lambda(-A))$. К сожалению, мы заранее не всегда будем знать, какая из матриц A или B плохо определена, хотя в некоторых задачах статистической или теоретической физики хорошая обусловленность B почти гарантируется.

В плохо обусловленном случае метод § 68 имеет некоторое преимущество в том, что вся плохая обусловленность B концентрируется в малых элементах D . Матрица P из (68.5) имеет определенное количество строк и столбцов с большими элементами (соответствующими малым d_{ii}), и собственные значения матрицы $(A - \lambda B)$ нормальной величины вероятнее всего должны сохраняться.

ПРИВЕДЕНИЕ ОБЩИХ МАТРИЦ К МАТРИЦАМ СПЕЦИАЛЬНОГО ВИДА

Введение

1. Рассмотрим теперь более трудную проблему вычисления системы собственных значений и собственных векторов матрицы общего вида A_0 . Мы не можем теперь ожидать *априорных* гарантий точности результатов, но мы можем ожидать, что существуют устойчивые преобразования, которые дают матрицы, в точности подобные $(A_0 + E)$, и что можно установить удовлетворительные границы для E .

В первой половине этой главы будем рассматривать приведение A_0 к форме Хессенберга (гл. 4, § 33) при помощи элементарных преобразований подобия. Несмотря на то, что, вообще говоря, матрицы Хессенберга имеют $\frac{1}{2}n(n+1)$ ненулевых элементов, т. е. в них аннулировано

только около половины элементов, приведение матрицы к форме Хессенберга вызывает значительное упрощение в решении проблемы собственных значений. Во всех методах, которые мы будем обсуждать, имеются две модификации, приводящие к верхней или нижней матрицам Хессенберга. С точки зрения дальнейших приложений несколько удобнее иметь матрицы в верхней форме, и потому будут описаны методы, соответствующие такой форме.

Это имеет одно неприятное следствие. Для симметричных матриц принято описывать методы в терминах элементов верхнего треугольника и, следовательно, получать нули в этом верхнем треугольнике, и мы следовали этому в главе 5. (Конечно, из-за симметрии нули одновременно получались и в нижнем треугольнике.) Наше решение поэтому делает сравнение методов для симметричных матриц с соответствующими методами для несимметричных матриц значительно менее простым, чем это могло бы быть, хотя модификации, необходимые для получения матриц в нижней форме Хессенберга, тривиальны.

Во второй половине главы мы рассмотрим дальнейшее приведение матриц Хессенберга к трехдиагональной форме и форме Фробениуса. Решение проблемы собственных значений для матриц этих специальных форм рассматривается в главе 7.

Метод Гивенса

2. Рассмотрим сначала аналоги методов Гивенса и Хаусхолдера, которые обсуждались в главе 5. Приведение по методу Гивенса состоит из $(n-2)$ основных шагов. Перед r -м шагом первые $(r-1)$ столбцов матрицы A_0 уже приведены к верхней форме Хессенберга, так что

например, при $n = 6$, $r = 3$ имеем

$$\begin{bmatrix} \times & \times & & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times \\ 0 & \times & & \times & \times & \times & \times \\ 0 & 0 & & \times & \times & \times & \times \\ 0 & 0 & \underline{\quad} & \times & \times & \times & \times \\ 0 & 0 & \underline{\quad} & \times & \times & \times & \times \end{bmatrix} \begin{matrix} \leftarrow \\ \leftarrow \\ \leftarrow \\ \leftarrow \\ \leftarrow \end{matrix} \quad (2.1)$$

$\uparrow \quad \uparrow \quad \uparrow$

Можно сказать, что главный ведущий минор порядка r уже приведен к форме Хессенберга, и эта часть не меняется при дальнейших шагах.

Основной r -й шаг состоит из $(n - r - 1)$ вспомогательных шагов, в которых последовательно аннулируются элементы в позициях $r + 2$, $r + 3$, ..., n r -го столбца. Ноль в i -й позиции получается при помощи вращения в плоскости $(r + 1, i)$. В (2.1) подчеркнуты элементы, которые нужно аннулировать, и отмечены стрелками строки и столбцы, которые при этом меняются. Основной r -й шаг может быть описан на общепонятном алгоритмическом языке следующим образом:

Для всех i от $r + 2$ до n выполнить шаги от (i) до (iv):

(i) вычислить $x = (a_{r+1, r}^2 + a_{ir}^2)^{1/2}$;

(ii) вычислить $\cos \theta = a_{r+1, r}/x$, $\sin \theta = a_{ir}/x$. (Если $x = 0$, взять $\cos \theta = 1$ и $\sin \theta = 0$ и опустить (iii) и (iv).) Записать x на место $a_{r+1, r}$ и ноль на место a_{ir} ;

(iii) для всех значений j от $r + 1$ до n вычислить $a_{r+1, j} \cos \theta + a_{ij} \sin \theta$ и $-a_{r+1, j} \sin \theta + a_{ij} \cos \theta$ и записать на место $a_{r+1, j}$ и a_{ij} соответственно;

(iv) для всех значений j от 1 до n вычислить $a_{j, r+1} \cos \theta + a_{ji} \sin \theta$ и $-a_{j, r+1} \sin \theta + a_{ji} \cos \theta$ и записать на место $a_{j, r+1}$ и место a_{ji} соответственно.

Шаг (iii) описывает эффект умножения слева, а шаг (iv) — умножения справа на вращение в плоскости $(r + 1, i)$. Это преобразование можно сравнить с соответствующим симметричным случаем, описанным в гл. 5, § 22. Из-за отсутствия симметрии требуется преобразование и строк и столбцов. Полное количество умножений приблизительно равно

$$\sum 4n(n - r - 1) + \sum 4(n - r - 1)^2 \div 2n^3 + \frac{4}{3}n^3 = \frac{10}{3}n^3, \quad (2.2)$$

тогда как в симметричном случае метода Гивенса $\frac{4}{3}n^3$. Заметим, что теперь нет места для запоминания величин $\cos \theta$ и $\sin \theta$ в позициях, в которых получаются нули, так как получение нуля в позиции (i, r) не связано с получением нуля в позиции (r, i) . Экономный метод запоминания, связанный с численной устойчивостью, состоит в следующем. Из значений $\cos \theta$ и $\sin \theta$ запоминается то, абсолютная величина которого меньше, а два последних разряда используются для указания того, был ли записан $\cos \theta$, и для знака незаписанной величины.

Метод Хаусхолдера

3. Подобным же образом имеем естественную аналогию метода Хаусхолдера, описанного в главе 5. Снова здесь $(n - 2)$ основных шага, и непосредственно перед r -м шагом A_0 имеет, например, для $n = 6, r = 3$, следующую форму:

$$A_{r-1} = \begin{matrix} & r & \left\{ \begin{array}{ccc|ccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ \hline 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \end{array} \right. \\ n-r & \end{matrix} = \left[\begin{array}{ccc|ccc} H_{r-1} & & & C_{r-1} & & \\ \hline 0 & b_{r-1} & & B_{r-1} & & \end{array} \right]. \quad (3.4)$$

Матрица H_{r-1} имеет верхнюю форму Хессенберга. Это аналогично форме, соответствующей началу r -го шага в методе Гивенса. Матрица A_{r-1} разделена так, чтобы проиллюстрировать эффект r -го преобразования. Матрица P_r этого преобразования может быть записана в виде

$$P_r = \begin{matrix} r \\ n-r \end{matrix} \left\{ \left[\begin{array}{c|c} I & 0 \\ \hline 0 & Q_r \end{array} \right] = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2v_r v_r^T \end{array} \right] \right\}, \quad (3.2)$$

где v_r — единичный вектор с $(n - r)$ компонентами. Следовательно, имеем

$$A_r = P_r A_{r-1} P_r = \begin{matrix} r \\ n-r \end{matrix} \left\{ \left[\begin{array}{c|c} H_{r-1} & C_{r-1} Q_r \\ \hline 0 & c_r \quad Q_r B_{r-1} Q_r \end{array} \right] \right\}, \quad (3.3)$$

где $c_r = Q_r b_{r-1}$. Если v_r выбрать так, чтобы c_r имел нулевые компоненты, за исключением первой, то тогда ведущая главная подматрица A_r порядка $(r + 1)$ будет в форме Хессенберга.

Снова, так же как в гл. 5, § 29, все формулы будут иметь простейший вид, если $P_r = I - 2w_r w_r^T$, где w_r — единичный вектор, имеющий n компонент, из которых первые r равны нулю. Тогда мы имеем

$$P_r = I - 2w_r w_r^T = I - u_r u_r^T / 2K_r^2, \quad (3.4)$$

где

$$\left. \begin{aligned} u_{ir} &= 0 & (i = 1, \dots, r), \\ u_{r+1, r} &= a_{r+1, r} \mp S_r, & u_{ir} = a_{ir} \quad (i = r + 2, \dots, n), \\ S_r &= \left(\sum_{i=r+1}^n a_{ir}^2 \right)^{1/2}, & 2K_r^2 = S_r^2 \mp a_{r+1, r} S_r. \end{aligned} \right\} \quad (3.5)$$

В (3.5) a_{ij} обозначает (i, j) -й элемент A_{r-1} . Новый $(r + 1, r)$ -й элемент равен $\pm S_r$. Знак выбирается обычным образом.

4. Так как теперь нет симметрии, умножения справа и слева на P_r нужно рассматривать отдельно. При умножении слева имеем

$$P_r A_{r-1} = (I - u_r u_r^T / 2K_r^2) A_{r-1} = A_{r-1} - u_r (u_r^T A_{r-1}) / 2K_r^2 = F_r, \quad (4.1)$$

а при умножении справа —

$$A_r = P_r A_{r-1} P_r = F_r P_r = F_r (I - u_r u_r^T / 2K_r^2) = F_r - (F_r u_r) u_r^T / 2K_r^2. \quad (4.2)$$

В силу того, что первые r элементов u_r равны нулю, из этих формул ясно, что умножение слева не меняет первые r строк A_{r-1} , а умножение справа получившейся F_r на P_r не меняет ее первые r столбцов.

Можно написать

$$u_r^T A_{r-1} = p_r^T, \quad (4.3)$$

где первые $(r-1)$ элементов p_r^T равны нулю из-за нулей u_r и A_{r-1} . Тогда имеем

$$F_r = A_{r-1} - (u_r/2K_r^2) p_r^T. \quad (4.4)$$

При этом не нужно вычислять r -й элемент p_r , так как от него зависит только r -й столбец F_r , а известно, что первые r элементов этого столбца совпадают с соответствующими элементами A_{r-1} , а остальные образуют вектор s_r . Преобразование было выбрано так, что первый элемент s_r равен $\pm S_r$, а остальные нули. Следовательно, требуется только $(n-r)^2$ умножений для вычисления $(n-r)$ элементов p_r и еще $(n-r)^2$ умножений при вычислении F_r по (4.4).

Для умножения справа положим

$$F_r u_r = q_r. \quad (4.5)$$

Вектор q_r не имеет нулевых компонент. Так как первые r компонент u_r равны нулю, при вычислении q_r требуется $n(n-r)$ умножений. Окончательно имеем

$$A_r = F_r - q_r (u_r/2K_r^2)^T, \quad (4.6)$$

и это требует еще $n(n-r)$ умножений. Заметим, что в (4.4) и (4.6) требуется вектор $u_r/2K_r^2$, а сам u_r необходим лишь для вычисления $F_r u_r$, получаемого последовательно по мере вычисления F_r . Эту трудность можно обойти, если работать с w_r , но ценой извлечения квадратного корня из $2K_r^2$.

Если матрица размещена во внешней памяти и главным требованием является экономия переносов, то следующая формулировка, вероятно, наиболее удобна. Определим векторы p_r^T , q_r , v_r и скаляр α_r соотношениями

$$p_r^T = u_r^T A_{r-1}, \quad q_r = A_{r-1} u_r, \quad v_r = u_r/2K_r^2, \quad p_r^T v_r = \alpha_r. \quad (4.7)$$

Затем вычислим A_r из соотношений

$$\begin{aligned} A_r &= A_{r-1} - v_r p_r^T - q_r v_r^T + (p_r^T v_r) u_r u_r^T = \\ &= A_{r-1} - v_r p_r^T - (q_r - \alpha_r u_r) v_r^T. \end{aligned} \quad (4.8)$$

Здесь много общего с симметричным случаем гл. 5, §§ 29 и 30, но член $u_r (u_r^T A_{r-1} u_r) u_r^T / (2K_r^2)^2$ теперь нельзя разделить между двумя другими членами.

Число умножений для полного приведения, с использованием любого из этих вариантов, равно приблизительно

$$\sum 2(n-r)^2 + \sum 2n(n-r) \approx \frac{2}{3} n^3 + n^3 = \frac{5}{3} n^3. \quad (4.9)$$

В симметричном варианте метода Хаусхолдера число умножений равно $\frac{2}{3} n^3$.

Снова замечаем, что метод Хаусхолдера требует вдвое меньше умножений, чем метод Гивенса.

Соображения о необходимой величине памяти

5. Запоминание параметров преобразования Хаусхолдера можно осуществить довольно удобным образом. Вектор u_r имеет $(n - r)$ ненулевых элементов, а соответствующее преобразование дает только $(n - r - 1)$ нулевых элементов. Следовательно, мы не можем хранить всю информацию на месте исходной матрицы. Однако мы увидим, что поддиагональные элементы матрицы Хессенберга играют специальную роль во всех дальнейших операциях, так что эти элементы удобно хранить отдельно в виде линейного массива из $(n - 1)$ элементов. Заметим, что $u_{i,r} = a_{i,r}$ ($i = r + 2, \dots, n$) и, следовательно, все ненулевые элементы $u_{i,r}$, кроме $u_{r+1,r}$, уже хранятся в соответствующих местах. Размещение элементов в конце r -го основного шага в случае $n = 6, r = 3$ имеет вид

$$\begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} & a_{15} & a_{16} \\ u_{21} & h_{22} & h_{23} & h_{24} & a_{25} & a_{26} \\ u_{31} & u_{32} & h_{33} & h_{34} & a_{35} & a_{36} \\ u_{41} & u_{42} & u_{43} & h_{44} & a_{45} & a_{46} \\ u_{51} & u_{52} & u_{53} & a_{54} & a_{55} & a_{56} \\ u_{61} & u_{62} & u_{63} & a_{64} & a_{65} & a_{66} \end{bmatrix}. \quad (5.1)$$

Здесь h_{ij} — элементы конечной матрицы Хессенберга, a_{ij} — текущие элементы A_r . В дополнение к этому мы используем n ячеек памяти для размещения вектора

$$[h_{21}, h_{32}, \dots, h_{r+1,r}; (u_{r+1,r}, u_{r+2,r}, \dots, u_{n,r})/2K_r^2]. \quad (5.2)$$

На вычислительных машинах с двухступенчатой памятью требуется еще группа из n рабочих ячеек для размещения вектора p_r или вектора q_r , ибо эти векторы не требуются одновременно. При одноступенчатой памяти нет надобности запоминать эти векторы, так как каждый элемент p_r и q_r может быть использован в соотношениях (4.4) и (4.6) немедленно после того, как он вычислен. При вычислении S_r^2 , p_r^T и q_r можно накапливать скалярные произведения.

Анализ ошибок

6. Общий метод главы 3 позволяет нам показать, что оба метода — Гивенса и Хаусхолдера — дают окончательно вычисленную матрицу Хессенберга, которая в точности подобна $(A + E)$, и дать оценки для E . Рассмотрение для несимметричного случая не труднее, чем для симметричного; наоборот, часто оно несколько проще. Полученные результаты приведены в табл. 1. Во всех случаях матрица преобразования подобия, приводящая $A + E$ к H , является точным произведением точных ортогональных матриц, соответствующих вычисленным A_r . В табл. 1 K_r суть величины порядка единицы. При получении первых двух оценок использование $\bar{f}_i$ вместо $\bar{f}_{i2}$ влияет только на постоянные K_1 и K_2 .

При вычислениях с фиксированной запятой требуется некоторое предварительное масштабирование для того, чтобы быть уверенными, что все промежуточные величины будут в допустимой области. Мы дали оценки для $\|E\|_2$ и $\|E\|_E$ там, где они были определены. В соответствии с оценкой $\|E\|_E$ мы дали оценку для возможной величины $\|A_0\|_E$, хотя,

Т а б л и ц а 1

Метод	Способ вычисления	Предварительное масштабирование	Границы для $\ E\ $
Гивенса	Фиксированная запятая (fi_2)	$\ A_0\ _2 \leq 1 - K_1 2^{-l} n^{5/2}$	$\ E\ _2 \leq K_1 2^{-l} n^{5/2}$
Гивенса	Фиксированная запятая (fi_2)	$\ A_0\ _E \leq 1 - K_2 2^{-l} n^{5/2}$	$\ E\ _E \leq K_2 2^{-l} n^{5/2}$
Хаусхолдера	Фиксированная запятая (fi)	$\ A_0\ _2 \leq \frac{1}{2} - K_3 2^{-l} n^{5/2}$	$\ E\ _2 \leq K_3 2^{-l} n^{5/2}$
Хаусхолдера	Фиксированная запятая (fi)	$\ A_0\ _E \leq \frac{1}{2} - K_4 2^{-l} n^{5/2}$	$\ E\ _E \leq K_4 2^{-l} n^{5/2}$
Хаусхолдера	Фиксированная запятая (fi_2)	$\ A_0\ _2 \leq \frac{1}{2} - K_5 2^{-l} n^2$	$\ E\ _2 \leq K_5 2^{-l} n^2$
Гивенса	Плавающая запя- тая (fl)	Не нужно	$\ E\ _E \leq K_6 2^{-l} n^{3/2} \ A_0\ _E$
Хаусхолдера	Плавающая запя- тая (fl)	Не нужно	$\ E\ _E \leq K_7 2^{-l} n^2 \ A_0\ _E$
Хаусхолдера	Плавающая запя- тая (fl_2)	Не нужно	$\ E\ _E \leq K_8 2^{-l} n \ A_0\ _E$

конечно, менее ограничительная оценка $\|A_0\|_2$ была бы достаточна для того, чтобы избежать превышения допустимых размеров. Однако оценку для $\|A_0\|_E$ легко получить на практике в отличие от оценки для $\|A_0\|_2$.

При использовании арифметики с плавающей запятой все оценки $\|E\|_E$ должны содержать множитель вида $(1 + f(n)2^{-l})^{g(n)}$, который стремится к бесконечности вместе с n , но так как этот множитель близок к единице там, где применение метода оправдано, пренебрежение им вполне допустимо.

Эти результаты показывают, что если A достаточно хорошо обусловлена, то ограничением для применения любого метода из табл. 1 является в большей мере емкость памяти и время, а не точность. Конечно, мы можем построить настолько плохо обусловленную A , что даже очень малая граница для E не обеспечит требуемую точность для собственных значений.

Связь методов Гивенса и Хаусхолдера

7. Как метод Гивенса, так и метод Хаусхолдера дают приведение A_0 к верхней матрице Хессенберга при помощи ортогонального преобразования подобия. Покажем, что матрицы, полученные при помощи этих двух преобразований, совпадают с точностью до множителя ± 1 в каждом столбце. Сначала докажем следующую несколько более общую теорему, которая понадобится нам впоследствии.

Если $AQ_1 = Q_1H_1$ и $AQ_2 = Q_2H_2$, где Q_i — унитарные, а H_i — верхние матрицы Хессенберга, и если первый столбец Q_1 совпадает с первым столбцом Q_2 , то, вообще говоря,

$$Q_2 = Q_1 D, \quad H_2 = D^H H_1 D, \quad (7.1)$$

где D — диагональная матрица, элементы которой по модулю равны единице.

Нам надо только показать, что если

$$AQ = QH \quad (7.2)$$

и первый столбец Q задан, то Q и H в основном однозначно определены. Доказательство следующее.

Приравнивая первые столбцы (7.2), имеем

$$Aq_1 = h_{11}q_1 + h_{21}q_2. \quad (7.3)$$

Следовательно, в силу унитарности Q

$$h_{11} = q_1^H Aq_1, \quad h_{21}q_2 = Aq_1 - h_{11}q_1. \quad (7.4)$$

Первое уравнение дает h_{11} , а второе

$$|h_{21}| = \|Aq_1 - h_{11}q_1\|_2, \quad (7.5)$$

так что q_2 определен с точностью до множителя $e^{i\theta_2}$.

Приравнивая вторые столбцы (7.2), имеем

$$Aq_2 = h_{12}q_1 + h_{22}q_2 + h_{32}q_3, \quad (7.6)$$

и следовательно,

$$h_{12} = q_1^H Aq_2, \quad h_{22} = q_2^H Aq_2, \quad h_{32}q_3 = Aq_2 - h_{12}q_1 - h_{22}q_2. \quad (7.7)$$

Первые два уравнения дают h_{12} и h_{22} , а третье —

$$|h_{32}| = \|Aq_2 - h_{12}q_1 - h_{22}q_2\|_2. \quad (7.8)$$

Можно легко проверить, что $|h_{32}|$ не зависит от множителя $e^{i\theta_2}$, так что q_3 определен с точностью до множителя $e^{i\theta_3}$. Продолжая таким образом, мы определим Q с точностью до умножения справа на множитель D , равный $\text{diag}(e^{i\theta_j})$, и очевидно, что соответствующая неопределенность H такая, как показано в (7.1).

Заметим, что если мы потребуем, чтобы $h_{i+1,i}$ были вещественными и положительными, то Q и H единственным образом определяются первым столбцом Q . Доказательство не проходит, если какой-либо из $h_{i+1,i}$ равен нулю, так как тогда q_{i+1} есть произвольный нормированный вектор, ортогональный $q_1, q_2, \dots, q_i$. Теперь эквивалентность методов Гивенса и Хаусхолдера следует, если мы заметим, что первый столбец матрицы каждого из преобразований равен e_1 . Это вытекает из того, что в методе Гивенса нет вращений, включающих первую плоскость, и все векторы w_r в методе Хаусхолдера имеют равный нулю первый элемент. Эквивалентность симметричных методов Гивенса и Хаусхолдера является, конечно, частным случаем только что доказанного результата. Доказательство следует сравнить с доказательством § 53 гл. 4. Снова мы подчеркиваем, что хотя матрицы полного преобразования Гивенса и Хаусхолдера в основном одинаковы, матрицы, при помощи которых получаются нули в r -м столбце, различны для этих двух методов. Доказательство почти совпадает с соответствующим доказательством, данным в гл. 4, § 53.

Несмотря на формальную эквивалентность этих двух преобразований и на тот факт, что оба они «устойчивы», на практике ошибки округления могут привести к различным окончательным результатам. Это не удивительно, так как мы видели в гл. 5, § 28, что даже если применить два раза к одной и той же матрице метод Гивенса, работая со слегка разной точностью, то можно получить сильно отличающиеся трехдиагональные матрицы. Мы видели, что такие расхождения двух вычислений должны иметь место, если на каком-либо основном шаге все элементы, которые нужно аннулировать, малы. Заметим, что согласно (7.5) и (7.8) элемент

$h_{i+1, i}$ равен квадратному корню из суммы квадратов таких элементов, и мы выше замечали, что соотношение $AQ = QH$ не определяет Q единственным образом, если $h_{i+1, i}$ равен нулю. Неединственность точного преобразования при $h_{i+1, i}$, равном нулю, вызывает плохую определенность преобразования на практике, если $h_{i+1, i}$ мал. Еще раз подчеркнем, что точность конечного результата никоим образом не зависит от плохой определенности преобразования.

Элементарные устойчивые преобразования

8. Преимущества использования ортогональных преобразований для симметричных матриц настолько велики, что мы не рассматривали использования более простых элементарных устойчивых матриц. Теперь обстоятельства сильно изменились.

Рассмотрим приведение A_0 к верхней форме Хессенберга при помощи устойчивых элементарных матриц типа N_r (гл. 3, § 47). Приведение состоит из $(n - 2)$ основных шагов. Перед началом r -го шага матрица A_0 приведена к A_{r-1} , которая для $n = 6$, $r = 3$ имеет вид

$$\begin{bmatrix} h_{11} & h_{12} & h_{13} & a_{14} & a_{15} & a_{16} \\ h_{21} & h_{22} & h_{23} & a_{24} & a_{25} & a_{26} \\ 0 & h_{32} & h_{33} & a_{34} & a_{35} & a_{36} \\ 0 & 0 & a_{43} & a_{44} & a_{45} & a_{46} \\ 0 & 0 & a_{53} & a_{54} & a_{55} & a_{56} \\ 0 & 0 & a_{63} & a_{64} & a_{65} & a_{66} \end{bmatrix}. \quad (8.1)$$

Мы увидим, что последующие шаги не меняют элементов, обозначенных h_{ij} . Основной r -й шаг состоит в следующем:

(i) определяем наибольшую из величин $|a_{ir}|$ ($i = r + 1, \dots, n$). (Если элементов с наибольшим модулем несколько, то берем первый из них.) Если этот элемент есть $a_{(r+1)', r}$, то переставляем строки $r + 1$ и $(r + 1)'$, т. е. умножаем слева на $I_{r+1, (r+1)'}$. (Элемент $a_{(r+1)', r}$ можно рассматривать как «ведущий элемент» по отношению к нашему преобразованию);

(ii) для всех значений i от $r + 2$ до n вычисляем $n_{i, r+1} = a_{ir}/a_{r+1, r}$, так что $|n_{i, r+1}| \leq 1$, и вычитаем из i -й строки $n_{i, r+1} \times$ (строка $r + 1$). Операции в (ii) означают умножение слева на матрицу типа N_{r+1} . С их помощью получаются нули в позициях (i, r) ($i = r + 2, \dots, n$);

(iii) переставляем столбцы $r + 1$ и $(r + 1)'$, т. е. умножаем справа на $I_{r+1, (r+1)'}$;

(iv) для всех значений i от $r + 2$ до n прибавляем к $r + 1$ -му столбцу $n_{i, r+1} \times$ (i -й столбец). Операции в (iv) означают умножение справа на N_{r+1}^{-1} . Если ведущий элемент равен нулю, то $n_{i, r+1}$ произвольны и наиболее просто взять $N_{r+1} I_{r+1, (r+1)'} = I$.

Заметим, что столбцы от первого до $(r - 1)$ -го не меняются при r -м преобразовании, которое само определяется элементами r -го столбца в позициях от $r + 1$ до n . Перестановки включены для того, чтобы избежать численной неустойчивости. Элементы r -го столбца от 1 до r также не меняются. Имеются $(n - r - 1)$ величин $n_{i, r+1}$, и их можно записать в позициях, которые раньше занимали аннулированные a_{ir} .

9. Кажется, что описание, данное в § 8, является наиболее естественным путем введения алгоритма. Но сейчас мы опишем две различные

модификации r -го основного шага, которые дадут возможность получать преимущества от накопления скалярных произведений. Мы снова будем использовать верхний индекс r , так как такие обозначения будут нужны при анализе ошибок в этих модификациях. Перестановки и множители $n_{i, r+1}$ определяются так же, как и раньше, и преобразованные элементы определяются из следующих соотношений.

Модификация I:

$$\left. \begin{aligned} a_{i, r+1}^{(r)} &= a_{i, r+1}^{(r-1)} + \sum_{j=r+2}^n a_{ij}^{(r-1)} n_{j, r+1} & (i = 1, \dots, r+1), \\ a_{i, r+1}^{(r)} &= a_{i, r+1}^{(r-1)} + \sum_{j=r+2}^n a_{ij}^{(r-1)} n_{j, r+1} - n_{i, r+1} a_{r+1, r+1}^{(r)} & (i = r+2, \dots, n), \\ a_{ij}^{(r)} &= a_{ij}^{(r-1)} - n_{i, r+1} a_{r+1, j}^{(r-1)} & (i, j = r+2, \dots, n). \end{aligned} \right\} \quad (9.1)$$

Модификация II:

$$\left. \begin{aligned} a_{ij}^{(r)} &= a_{ij}^{(r-1)} - n_{i, r+1} a_{r+1, j}^{(r-1)} & (i, j = r+2, \dots, n), \\ a_{i, r+1}^{(r)} &= a_{i, r+1}^{(r-1)} - n_{i, r+1} a_{r+1, r+1}^{(r-1)} + \sum_{j=r+2}^n a_{ij}^{(r-1)} n_{j, r+1} & (i = r+2, \dots, n), \\ a_{i, r+1}^{(r)} &= a_{i, r+1}^{(r-1)} + \sum_{j=r+2}^n a_{ij}^{(r-1)} n_{j, r+1} & (i = 1, \dots, r+1). \end{aligned} \right\} \quad (9.2)$$

Единственная ошибка округления делается тогда, когда заканчивается накопление соответствующих скалярных произведений. Эти два процесса математически идентичны, но различны на практике. Так как ведущий элемент следующего шага выбирается из элементов $a_{i, r+1}^{(r)}$ ($i = r+2, \dots, n$), мы можем определять $(r+2)'$ сразу, как только эти элементы вычислены.

Приблизительное число умножений в r -м шаге равно $(n-r)^2$ при умножении слева и $n(n-r)$ — при умножении справа, независимо от того, используем ли мы исходную формулировку или модификации. Общее число умножений в полном приведении приблизительно равно

$$\sum (n-r)^2 + \sum n(n-r) \approx \frac{5}{6} n^3, \quad (9.3)$$

что равно половине числа умножений при приведении по методу Хаусхолдера и четверти — по методу Гивенса.

Значение перестановок

10. Полное преобразование можно записать в виде

$$N_{n-1} I_{n-1, (n-1)} \dots N_3 I_{3, 3} \cdot N_2 I_{2, 2} \cdot A_0 I_{2, 2} \cdot N_2^{-1} I_{3, 3} \cdot N_3^{-1} \dots I_{n-1, (n-1)} \cdot N_{n-1}^{-1} = H, \quad (10.1)$$

где H — окончательная матрица в форме Хессенберга. Покажем, что перестановки определяют такое подобное преобразование A_0 , что ведущие элементы возникают естественным образом в нужных позициях.

Мы ограничимся доказательством для случая матриц пятого порядка. Общий случай очевидным образом следует из него (см. гл. 4, § 22). При $n = 5$ уравнение (10.1) может быть записано в виде

$$N_4 I_{4,4} N_3 (I_{4,4} I_{4,4}) I_{3,3} N_2 (I_{3,3} I_{4,4} I_{4,4} I_{3,3}) I_{2,2} A_0 \times \\ \times I_{2,2} (I_{3,3} I_{4,4} I_{4,4} I_{3,3}) N_2^{-1} I_{3,3} (I_{4,4} I_{4,4}) N_3^{-1} I_{4,4} N_4^{-1} = H, \quad (10.2)$$

так как каждое произведение в скобках равно I . Группируя сомножители, получаем

$$(N_4) (I_{4,4} N_3 I_{4,4}) (I_{4,4} I_{3,3} N_2 I_{3,3} I_{4,4}) (I_{4,4} I_{3,3} I_{2,2} A_0 I_{2,2} I_{3,3} I_{4,4}) \times \\ \times (I_{4,4} I_{3,3} N_2^{-1} I_{3,3} I_{4,4}) (I_{4,4} N_3^{-1} I_{4,4}) N_4^{-1} = H. \quad (10.3)$$

Матрица в середине является подобным преобразованием A_0 с матрицей перестановок в качестве трансформирующей. Мы можем назвать ее $\tilde{A}_0$. Если мы обозначим множители в первых трех скобках слева $\tilde{N}_4, \tilde{N}_3, \tilde{N}_2$, то (10.3) даст

$$\tilde{N}_4 \tilde{N}_3 \tilde{N}_2 \tilde{A}_0 \tilde{N}_2^{-1} \tilde{N}_3^{-1} \tilde{N}_4^{-1} = H. \quad (10.4)$$

Так как $r' \geq r$ ($r = 2, \dots, n-1$), каждая $\tilde{N}_r$ является матрицей в точности того же вида, что N_r , только поддиагональные элементы в r -м столбце переставлены (гл. 1, § 41 (iv)). Следовательно, $\tilde{A}_0$, получаемая из A_0 одновременной перестановкой строк и столбцов, может быть приведена к H без перестановок, и множители автоматически будут по модулю ограничены единицей. Из гл. 1, § 41 (iii) известно, что в матрице $\tilde{N}_2^{-1} \tilde{N}_3^{-1} \tilde{N}_4^{-1}$ нет произведений n_{ij} . Мы можем обозначить это произведение $\tilde{N}$, где $\tilde{N}$ — единичная треугольная матрица, первый столбец которой равен e_1 , так как в произведении нет сомножителя $\tilde{N}_1^{-1}$. Элементы r -го столбца матрицы $\tilde{N}$ под диагональю равны соответствующим элементам переставленным элементам n_{ir} . Итак, мы имеем

$$\tilde{A}_0 \tilde{N} = \tilde{N} H, \quad (10.5)$$

и если при преобразовании A_0 не нужны перестановки, это даст

$$A_0 N = N H. \quad (10.6)$$

В гл. 4, § 5 мы указывали на необходимость уравнивания матрицы перед выполнением метода Гаусса с выбором главного элемента. Подобным же образом в нашем случае трудно выбирать ведущие элементы, если только не выполнено некоторого аналога уравнивания. В случае проблемы собственных значений это может быть достигнуто, например, при помощи замены A_0 на DA_0D^{-1} , где D — диагональная матрица, выбранная так, чтобы наибольшие по модулю элементы каждой строки и каждого столбца преобразованной матрицы были одного порядка. До сих пор не проведено достаточно полного анализа этой проблемы. На АСЕ мы испытывали разные формы уравнивания. Нет никаких сомнений в том, что очень важно избегать работы с плохо сбалансированными матрицами. (Обсуждение этой проблемы см. Бауэр (1963) и Осборн (1960).)

Прямое приведение к форме Хессенберга

11. В гл. 4, § 17 мы видели, что метод Гаусса без выбора главного элемента дает такие треугольные матрицы N и A_{n-1} , что

$$NA_{n-1} = A_0. \quad (11.1)$$

Затем в § 36 мы показали, что элементы N и A_{n-1} могут определяться приравниванием элементов в обеих частях (11.1). Далее было показано, как может быть осуществлен выбор ведущего элемента. Покажем аналогично, на время игнорируя выбор ведущего элемента, что уравнение (10.6) можно использовать для нахождения непосредственно элементов N и H , минуя все промежуточные стадии определения A_r . Соответственно опустим нулевой индекс у A_0 , так как он уже не имеет отношения к делу.

При $n = 5$ уравнение (10.6) имеет вид

$$\begin{array}{c} A \qquad \qquad N \\ \left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right] \left[\begin{array}{ccccc} 1 & & & & \\ 0 & 1 & & & \\ 0 & \times & 1 & & \\ 0 & \times & \times & 1 & \\ 0 & \times & \times & \times & 1 \end{array} \right] = \\ \\ \qquad \qquad N \qquad \qquad H \\ = \left[\begin{array}{ccccc} 1 & & & & \\ 0 & 1 & & & \\ 0 & \times & 1 & & \\ 0 & \times & \times & 1 & \\ 0 & \times & \times & \times & 1 \end{array} \right] \left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \end{array} \right]. \quad (11.2) \end{array}$$

Используя структуры этих матриц, покажем, что, приравнивая последовательно столбцы в обеих частях, можно определить все элементы N и H столбец за столбцом. На самом деле, приравнивая r -е столбцы, определяем r -й столбец H и $(r+1)$ -й столбец N . Как *заранее* известно, первый столбец N равен e_1 .

Доказательство по индукции. Предположим, что приравнивая первые $(r-1)$ столбцов в уравнении (10.6), мы определили первые $(r-1)$ столбцов H и первые r столбцов N . Приравняем элементы r -го столбца. Тогда (i, r) -е элементы в левой стороне равны

$$a_{ir} + a_{i, r+1}n_{r+1, r} + a_{i, r+2}n_{r+2, r} + \dots + a_{in}n_{n, r} \quad (\text{для всех } i), \quad (11.3)$$

в то время как (i, r) -е элементы в правой стороне равны

$$\left. \begin{array}{l} n_{i1}h_{1r} + n_{i2}h_{2r} + \dots + n_{i, i-1}h_{i-1, r} + h_{ir} \quad (i \leq r+1), \\ n_{i1}h_{1r} + n_{i2}h_{2r} + \dots + n_{i, r+1}h_{r+1, r} \quad (i > r+1), \end{array} \right\} \quad (11.4)$$

где мы включили члены с n_{i1} , которые равны нулю, для того, чтобы первые слагаемые не были особыми. Так как r -й столбец N уже известен, выражение (11.3) можно вычислить при всех значениях i . Приравнивая элементы, последовательно получаем h_{1r} , h_{2r} , $\dots$, $h_{r+1, r}$; $n_{r+2, r+1}$, $n_{r+3, r+1}$, $\dots$, $n_{n, r+1}$. Это отличные от нуля элементы r -го

столбца H и $(r+1)$ -го столбца N . Процесс заканчивается, когда определится n -й столбец H . Трудностей в начале процесса не возникает, так как первый столбец N известен.

Эта формулировка приведения позволяет получить все преимущества накопления скалярных произведений. Так, из (11.3) и (11.4) имеем формулы

$$\left. \begin{aligned} h_{ir} &= a_{ir} \times 1 + \sum_{k=r+1}^n a_{ik} n_{kr} - \sum_{k=1}^{i-1} n_{ik} h_{kr} & (i=1, \dots, r+1), \\ n_{i, r+1} &= (a_{ir} \times 1 + \sum_{k=r+1}^n a_{ik} n_{kr} - \sum_{k=1}^n n_{ik} h_{kr}) / h_{r+1, r} & (i=r+2, \dots, n), \end{aligned} \right\} \quad (11.5)$$

и h_{ir} имеет вид скалярного произведения, а $n_{i, r+1}$ — скалярного произведения, деленного на $h_{r+1, r}$.

Заметим, что мы можем определить целиком N и H этим способом потому, что а priori известен первый столбец N . Не слишком важно, что этот столбец равен e_1 . Если взять в качестве первого столбца произвольный вектор, первый элемент которого равен единице, то мы снова сможем определить целиком N и H , и уравнения не изменятся. (Результат, который мы сейчас доказали, является близкой аналогией результата § 7.) Частично по этой причине мы включили члены n_{i1} ($i \neq 1$). Если мы следуем формулировке § 8, то выбор первого столбца N отличным от e_1 эквивалентен вычислению $N_1 A_0 N_1^{-1}$ перед началом обычного процесса. Это замечание будет полезно позднее.

Введение перестановок

12. Процесс § 11 формально совпадает с процессом § 8, но без учета перестановок. Их использование весьма просто, и лучше всего описывается на алгоритмическом языке. Имеется n основных шагов; на r -м шаге мы определяем r -й столбец H и $(r+1)$ -й столбец N . Проиллюстрируем конфигурацию непосредственно перед началом r -го шага для случая $n=6$, $r=3$:

$$\begin{array}{cccccc} \left[\begin{array}{cccccc} h_{11} & h_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ h_{21} & h_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ n_{32} & h_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ n_{42} & n_{43} & a_{43} & a_{44} & a_{45} & a_{46} \\ n_{52} & n_{53} & a_{53} & a_{54} & a_{55} & a_{56} \\ n_{62} & n_{63} & a_{63} & a_{64} & a_{65} & a_{66} \end{array} \right] & \begin{array}{l} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{array} \\ r = & \begin{array}{cc} 2' & 3' \end{array} \end{array} \quad (12.1)$$

Здесь a_{ij} это элементы исходной матрицы A , хотя перестановки и могут иметь место; эти перестановки и значение рабочих ячеек σ_i и r' станут понятными, когда мы опишем r -й шаг.

(i) Для всех значений i от 1 до r вычисляем $h_{ir} = a_{ir} + \sum_{k=r+1}^n a_{ik} n_{kr} - \sum_{k=2}^{i-1} n_{ik} h_{kr}$, накапливая, если возможно, скалярные произведения. Окружаем окончательное значение до одинарной точности и записываем на место a_{ir} . (Если верхний предел какой-либо суммы меньше нижнего,

то этой суммой пренебрегаем. Помним, что $n_{i1} = 0$ ($i = 2, \dots, n$). Если $r = n$, приведение завершено.

(ii) Для всех значений i от $r+1$ до n вычисляем $a_{ir} + \sum_{k=r+1}^n a_{ik}n_{kr} - \sum_{k=2}^r n_{ik}h_{kr}$ и записываем в рабочей ячейке σ_i . Каждая сумма, если возможно, вычисляется с двойной точностью, и поэтому необходимы две ячейки.

После того как все σ_i вычислены, выявляем $\max_{i=r+1}^n |\sigma_i|$, и если $\max_{i=r+1}^n |\sigma_i| = |\sigma_{(r+1)'}|$, то запоминаем $(r+1)'$. Если максимум достигается несколькими σ_i , берем первую такую σ_i .

(iii) Переставляем строки $r+1$ и $(r+1)'$ (включая n_{ij} и σ_i) и столбцы $r+1$ и $(r+1)'$. Округляем текущее σ_{r+1} до одинарной точности, получаем h_{r+1} , r и записываем на место $a_{r+1, r}$. Это ведущий элемент.

(iv) Для всех значений i от $r+2$ до n вычисляем $n_{i, r+1} = \sigma_i/h_{r+1, r}$ и записываем на место a_{ir} .

Некоторые пункты требуют специального комментария. Если на каком-либо этапе

$$\max_{i=r+1}^n |\sigma_i| = 0,$$

то $n_{i, r+1}$ ($i = r+2, \dots, n$) не определены, и мы можем взять любые значения, по модулю меньшие единицы. При этом сохраняется численная устойчивость. Наиболее просто взять соответствующие $n_{i, r+1}$ равными нулю. Если используется арифметика с плавающей запятой и накоплением скалярных произведений, то мы мало теряем, если σ_i округляются до одинарной точности сразу после вычисления. Тогда каждое σ_i временно может храниться на месте соответствующего a_{ir} до тех пор, пока оно не понадобится для вычисления $n_{i, r+1}$. В этом случае мы можем обойтись без ячеек, обозначенных σ_i . При использовании накопления с фиксированной запятой мы должны запоминать σ_i с двойной точностью, если желательно получить полную выгоду. Однако, так как таких σ_i на r -м шаге только $n - r$, то в ячейках σ_i всегда есть место для величин $2', 3', \dots, r'$. Надо подчеркнуть, что, за исключением ошибок округления, процесс, который мы сейчас описали, совпадает с процессом в §§ 8, 9. При точных вычислениях перестановки h_{ij} и n_{ij} одинаковы.

Численный пример

13. В табл. 2 мы даем численный пример приведения матрицы пятого порядка при помощи метода § 12, используя арифметику с четырьмя десятичными знаками и фиксированной запятой с накоплением скалярных произведений. Пример представлен очень детально, так как алгоритм значительно сложнее, чем большинство описанных ранее. На вычислительных машинах процесс замечательно компактен, так как все данные размещаются в одном и том же объеме памяти величиной n^2 . Заметим, что из-за того, что первый столбец N_1 равен e_1 , первый шаг может быть несколько отличным от остальных. (Если мы модифицируем процедуру и позволим брать в качестве первого столбца произвольный вектор, первый элемент которого равен единице, а остальные ограничены по модулю единицей, то первый шаг перестанет отличаться от остальных.) Нужно быть особенно осторожным, программируя первый и последний

Таблица 2

А

$\left[\begin{array}{c} 0,3200 \\ 0,2500 \\ 0,4300 \\ 0,5300 \\ 0,1600 \end{array} \right]$	$\left[\begin{array}{c} 0,2700 \\ 0,0300 \\ 0,7300 \\ 0,2500 \\ 0,6500 \end{array} \right]$	$\left[\begin{array}{c} 0,2300 \\ 0,7100 \\ 0,1300 \\ 0,5100 \\ 0,4600 \end{array} \right]$	$\left[\begin{array}{c} 0,3200 \\ 0,2100 \\ 0,3700 \\ 0,6200 \\ 0,5600 \end{array} \right]$	$\left[\begin{array}{c} 0,4000 \\ 0,1700 \\ 0,8500 \\ 0,1600 \\ 0,3200 \end{array} \right]$	$\left. \begin{array}{l} h_{11} = 0,3200\ 0000 \\ \sigma_2 = 0,2500\ 0000 \\ \sigma_3 = 0,4300\ 0000 \\ \sigma_4 = 0,5300\ 0000 \\ \sigma_5 = 0,1600\ 0000 \end{array} \right\}$
$\max \sigma_i = \sigma_4 ; 2' = 4$					
<i>Конфигурация перед началом второго шага</i>					
$\left[\begin{array}{c} 0,3200 \\ 0,5300 \\ (0,8113) \\ (0,4717) \\ (0,3019) \end{array} \right]$	$\left[\begin{array}{c} 0,3200 \\ 0,6200 \\ 0,3700 \\ 0,2100 \\ 0,5600 \end{array} \right]$	$\left[\begin{array}{c} 0,2300 \\ 0,5100 \\ 0,1300 \\ 0,7100 \\ 0,4600 \end{array} \right]$	$\left[\begin{array}{c} 0,2700 \\ 0,2500 \\ 0,7300 \\ 0,0300 \\ 0,6500 \end{array} \right]$	$\left[\begin{array}{c} 0,4000 \\ 0,1600 \\ 0,8500 \\ 0,1700 \\ 0,3200 \end{array} \right]$	$\left. \begin{array}{l} h_{12} = 0,3200 + 0,2300 \times 0,8113 + 0,2700 \times 0,4717 + 0,4000 \times 0,3019 \\ h_{22} = 0,6200 + 0,5100 \times 0,8113 + 0,2500 \times 0,4717 + 0,1600 \times 0,3019 \\ \sigma_3 = 0,3700 + 0,1300 \times 0,8113 + 0,7300 \times 0,4717 + 0,8500 \times 0,3019 - 0,8113 \times 1,2000 = 0,1028\ 6500 \\ \sigma_4 = 0,2100 + 0,7100 \times 0,8113 + 0,0300 \times 0,4717 + 0,1700 \times 0,3019 - 0,4717 \times 1,2000 = 0,2854\ 5700 \\ \sigma_5 = 0,5600 + 0,4600 \times 0,8113 + 0,6500 \times 0,4717 + 0,3200 \times 0,3019 - 0,3019 \times 1,2000 = 0,9741\ 3100 \end{array} \right\}$
$\max \sigma_i = \sigma_5 ; 3' = 5$					

Заметим, что округленное значение h_{22} немедленно требуется для вычисления σ_i .

Конфигурация перед началом третьего шага

$\left[\begin{array}{c} 0,3200 \\ 0,5300 \\ (0,3019) \\ (0,4717) \\ (0,8113) \end{array} \right]$	$\left[\begin{array}{c} 0,7547 \\ 1,2000 \\ 0,9741 \\ (0,2930) \\ (0,4056) \end{array} \right]$	$\left[\begin{array}{c} 0,4000 \\ 0,1600 \\ 0,3200 \\ 0,1700 \\ 0,8500 \end{array} \right]$	$\left[\begin{array}{c} 0,2700 \\ 0,2500 \\ 0,6500 \\ 0,0300 \\ 0,7300 \end{array} \right]$	$\left[\begin{array}{c} 0,2300 \\ 0,5100 \\ 0,4600 \\ 0,7100 \\ 0,1300 \end{array} \right]$
--	--	--	--	--

$$\begin{aligned}
 h_{13} &= 0,4000 + 0,2700 \times 0,2930 + 0,2300 \times 0,1056 & = 0,5033 \ 9800 \\
 h_{23} &= 0,4600 + 0,2500 \times 0,2930 + 0,5100 \times 0,1056 & = 0,2871 \ 0600 \\
 h_{33} &= 0,3200 + 0,6500 \times 0,2930 + 0,4600 \times 0,1056 - 0,3019 \times 0,2871 & = 0,4723 \ 5051 \\
 \sigma_4 &= 0,1700 + 0,0300 \times 0,2930 + 0,7100 \times 0,1056 - 0,4717 \times 0,2871 - 0,2930 \times 0,4724 = -0,0200 \ 7227 \\
 \sigma_5 &= 0,8500 + 0,7300 \times 0,2930 + 0,1300 \times 0,1056 - 0,8113 \times 0,2871 - 0,1056 \times 0,4724 = 0,7948 \ 0833
 \end{aligned}
 \left. \vphantom{\begin{aligned} h_{13} \\ h_{23} \\ h_{33} \\ \sigma_4 \\ \sigma_5 \end{aligned}} \right\}$$

$$\max |\sigma_i| = |\sigma_5|; \quad 4' = 5.$$

Округленное значение h_{23} используется сразу не только при вычислении σ_i , но также и при вычислении h_{33} .

Конфигурация перед началом четвертого шага

$\left[\begin{array}{c} 0,3200 \\ 0,5300 \\ (0,3019) \\ (0,8113) \\ (0,4717) \end{array} \right]$	$\left[\begin{array}{c} 0,7547 \\ 1,2000 \\ 0,9741 \\ (0,1056) \\ (0,2930) \end{array} \right]$	$\left[\begin{array}{c} 0,5034 \\ 0,2871 \\ 0,4724 \\ 0,7948 \\ (-0,0253) \end{array} \right]$	$\left[\begin{array}{c} 0,2300 \\ 0,5100 \\ 0,4600 \\ 0,1300 \\ 0,7100 \end{array} \right]$	$\left[\begin{array}{c} 0,2700 \\ 0,2500 \\ 0,6500 \\ 0,7300 \\ 0,0300 \end{array} \right]$
--	--	---	--	--

$$\begin{aligned}
 h_{14} &= 0,2300 + (-0,0253) \times 0,2700 & = 0,2231 \ 6900 \\
 h_{24} &= 0,5100 + (-0,0253) \times 0,2500 & = 0,5036 \ 7500 \\
 h_{34} &= 0,4600 + (-0,0253) \times 0,6500 - 0,3019 \times 0,5037 & = 0,2914 \ 8797 \\
 h_{44} &= 0,1300 + (-0,0253) \times 0,7300 - 0,8113 \times 0,5037 - 0,1056 \times 0,2915 & = -0,3279 \ 0321 \\
 \sigma_5 &= 0,7100 + (-0,0253) \times 0,0300 - 0,4717 \times 0,5037 - 0,2930 \times 0,2915 - (-0,0253) \times (-0,3279) = 0,3779 \ 4034
 \end{aligned}$$

Конечно, σ_5 это h_{54} , так как сейчас есть лишь одна σ_i

Конфигурация перед началом пятого шага

$\left[\begin{array}{c} 0,3200 \\ 0,5300 \\ (0,3019) \\ (0,8113) \\ (0,4717) \end{array} \right]$	$\left[\begin{array}{c} 0,7547 \\ 1,2000 \\ 0,9741 \\ (0,1056) \\ (0,2930) \end{array} \right]$	$\left[\begin{array}{c} 0,5034 \\ 0,2871 \\ 0,4724 \\ 0,7948 \\ (-0,0253) \end{array} \right]$	$\left[\begin{array}{c} 0,2232 \\ 0,5037 \\ 0,2915 \\ -0,3279 \\ 0,3779 \end{array} \right]$	$\left[\begin{array}{c} 0,2700 \\ 0,2500 \\ 0,6500 \\ 0,7300 \\ 0,0300 \end{array} \right]$
--	--	---	---	--

$$\begin{aligned}
 h_{15} &= 0,2700 & = 0,2700 \ 0000 \\
 h_{25} &= 0,2500 & = 0,2500 \ 0000 \\
 h_{35} &= 0,6500 - 0,3019 \times 0,2500 & = 0,5745 \ 2500 \\
 h_{45} &= 0,7300 - 0,8113 \times 0,2500 - 0,1056 \times 0,5745 & = 0,4665 \ 0780 \\
 h_{55} &= 0,0300 - 0,4717 \times 0,2500 - 0,2930 \times 0,5745 - (-0,0253) \times 0,4665 = -0,2444 \ 5105
 \end{aligned}$$

Продолжение табл. 2

Окончательная конфигурация

$\begin{bmatrix} 0,3200 \\ 0,5300 \\ (0,3049) \\ (0,8113) \\ (0,4717) \end{bmatrix}$	$\begin{bmatrix} 0,7547 \\ 1,2000 \\ 0,9741 \\ (0,1056) \\ (0,2930) \end{bmatrix}$	$\begin{bmatrix} 0,5034 \\ 0,2871 \\ 0,4724 \\ 0,7948 \\ (-0,0253) \end{bmatrix}$	$\begin{bmatrix} 0,2232 \\ 0,5037 \\ 0,2915 \\ -0,3279 \\ 0,3779 \end{bmatrix}$	$\begin{bmatrix} 0,2700 \\ 0,2500 \\ 0,5745 \\ 0,4665 \\ -0,2445 \end{bmatrix}$
--	--	---	---	---

Отличные от нуля и единицы элементы матрицы N поставлены в скобкиПерестановка $\tilde{A}$ матрицы A , определенная преобразованием

$$\tilde{A} = I_{45} I_{35} I_{24} A I_{24} I_{35} I_{45}$$

$\begin{bmatrix} 0,3200 \\ 0,5300 \\ 0,1600 \\ 0,4300 \\ 0,2500 \end{bmatrix}$	$\begin{bmatrix} 0,3200 \\ 0,6200 \\ 0,5600 \\ 0,3700 \\ 0,2100 \end{bmatrix}$	$\begin{bmatrix} 0,4000 \\ 0,1600 \\ 0,3200 \\ 0,8500 \\ 0,1700 \end{bmatrix}$	$\begin{bmatrix} 0,2300 \\ 0,5100 \\ 0,4600 \\ 0,1300 \\ 0,7100 \end{bmatrix}$	$\begin{bmatrix} 0,2700 \\ 0,2500 \\ 0,6500 \\ 0,7300 \\ 0,0300 \end{bmatrix}$
--	--	--	--	--

$\tilde{A}N = \begin{bmatrix} 0,3200 \\ 0,5300 \\ 0,1600 \\ 0,4300 \\ 0,2500 \end{bmatrix}$	$\begin{bmatrix} 0,754718 \\ 1,199992 \\ 1,336411 \\ 1,076425 \\ 0,851497 \end{bmatrix}$	$\begin{bmatrix} 0,503398 \\ 0,287106 \\ 0,559026 \\ 1,077618 \\ 0,253766 \end{bmatrix}$	$\begin{bmatrix} 0,223169 \\ 0,503675 \\ 0,443555 \\ 0,111531 \\ 0,709241 \end{bmatrix}$	$\begin{bmatrix} 0,2700 \\ 0,2500 \\ 0,6500 \\ 0,7300 \\ 0,0300 \end{bmatrix}$
---	--	--	--	--

$NH = \begin{bmatrix} 0,3200 \\ 0,5300 \\ 0,160007 \\ 0,429989 \\ 0,250001 \end{bmatrix}$	$\begin{bmatrix} 0,7547 \\ 1,2000 \\ 1,336380 \\ 1,07642496 \\ 0,85145130 \end{bmatrix}$	$\begin{bmatrix} 0,5034 \\ 0,2871 \\ 0,55907549 \\ 1,07760967 \\ 0,25372983 \end{bmatrix}$	$\begin{bmatrix} 0,2232 \\ 0,5037 \\ 0,44356703 \\ 0,11153421 \\ 0,70920066 \end{bmatrix}$	$\begin{bmatrix} 0,2700 \\ 0,2500 \\ 0,64997500 \\ 0,72999220 \\ 0,02995105 \end{bmatrix}$
---	--	--	--	--

шаги этого алгоритма. Стрелки указывают переставляемые строки и столбцы. Подчеркнутые элементы — это элементы окончательной матрицы H , а элементы в скобках — это элементы N .

В конце примера мы привели матрицу $\tilde{A}$, которая дала бы окончательные матрицы N и H без использования перестановок. И, наконец, мы дали *точные* произведения *вычисленной* N и *вычисленной* H , и $\tilde{A}$ и *вычисленной* N . Они с точностью до ошибок округления должны совпадать. Сравнение этих двух произведений показывает, что они нигде не отличаются более чем на $\frac{1}{2} 10^{-4}$, максимальную величину одной ошибки округления.

Анализ ошибок

14. Покажем теперь, что замечание, сделанное относительно величины элементов $(\tilde{A}N - NH)$, верно в общем при некоторых предположениях.

Предположим, что выбор ведущего элемента производился способом, который был описан, что вычисления производились с использованием арифметики с фиксированной запятой и накоплением скалярных произведений и что все элементы A и вычисленной H меньше единицы. Тогда из (11.3) и (11.4) при $i \leq r + 1$ имеем

$$\begin{aligned} h_{ir} &= fi_2(a_{ir} + a_{i,r+1}n_{r+1,r} + \dots + a_{in}n_{nr} - n_{i1}h_{1r} - \dots - n_{i,i-1}h_{i-1,r}) \equiv \\ &\equiv a_{ir} + a_{i,r+1}n_{r+1,r} + \dots + a_{in}n_{nr} - n_{i1}h_{1r} - \dots - n_{i,i-1}h_{i-1,r} + \varepsilon_{ir}, \end{aligned} \quad (14.1)$$

где

$$|\varepsilon_{ir}| \leq \frac{1}{2} 2^{-t}. \quad (14.2)$$

Аналогично при $i > r + 1$ имеем

$$\begin{aligned} n_{i,r+1} &= fi_2[(a_{ir} + a_{i,r+1}n_{r+1,r} + \dots + a_{in}n_{nr} - n_{i1}h_{1r} - \dots - n_{ir}h_{rr})/h_{r+1,r}] \equiv \\ &\equiv (a_{ir} + a_{i,r+1}n_{r+1,r} + \dots + a_{in}n_{nr} - n_{i1}h_{1r} - \dots - n_{ir}h_{rr})/h_{r+1,r} + \eta_{ir}, \end{aligned} \quad (14.3)$$

где

$$|\eta_{ir}| \leq \frac{1}{2} 2^{-t}. \quad (14.4)$$

Умножая (14.3) на $h_{r+1,r}$, получим

$$\begin{aligned} n_{i,r+1}h_{r+1,r} &\equiv \\ &\equiv a_{ir} + a_{i,r+1}n_{r+1,r} + \dots + a_{in}a_{nr} - n_{i1}h_{1r} - \dots - n_{ir}h_{rr} + \varepsilon_{ir}, \end{aligned} \quad (14.5)$$

где

$$\varepsilon_{ir} = h_{r+1,r}\eta_{ir}. \quad (14.6)$$

Уравнения (14.1) и (14.5) показывают, что

$$(NH - \tilde{A}N)_{ir} = \varepsilon_{ir} \quad (i, r = 1, 2, \dots, n), \quad (14.7)$$

а из (14.2) и (14.6) следует, что

$$|\varepsilon_{ir}| \leq \frac{1}{2} 2^{-t} \quad (i, r = 1, 2, \dots, n), \quad (14.8)$$

так как мы предположили, что все элементы $|H|$ ограничены единицей. Следовательно, окончательно имеем

$$\tilde{A}N - NH = F, \quad (14.9)$$

где

$$|f_{ij}| \leq \frac{1}{2} 2^{-t}. \quad (14.10)$$

Это является вполне удовлетворительным результатом, так как вряд ли мы могли ожидать, что $(\tilde{A}N - NH)$ будет меньше! Однако он далеко не так хорош, как кажется. Уравнение (14.9) означает, что

$$H = N^{-1}(\tilde{A} - FN^{-1})N \quad (14.11)$$

и что

$$N^{-1}\tilde{A}N = H + N^{-1}F. \quad (14.12)$$

Следовательно, вычисленная H является точным преобразованием подобия $(\tilde{A} - FN^{-1})$ с вычисленной N в качестве матрицы преобразования. Другими словами, мы можем сказать, что H отличается от точного преобразования подобия $\tilde{A}$ на матрицу $N^{-1}F$. Тот факт, что вместо A привлечена $\tilde{A}$, несуществен, так как собственные значения и индивидуальные числа обусловленности у матриц A и $\tilde{A}$ совпадают.

Имеем

$$\|F\|_2 \leq \|F\|_E \leq \frac{1}{2} n 2^{-t}, \quad (14.13)$$

но хотя элементы n_{ij} ограничены единицей, $\|N^{-1}\|_2$ может быть весьма большой. В самом деле, если все соответствующие отличные от нуля n_{ij} равны -1 , то, например, для случая $n = 5$, мы имеем

$$N^{-1} = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 1 & 1 & & \\ 0 & 2 & 1 & 1 & \\ 0 & 4 & 2 & 1 & 1 \end{bmatrix}, \quad (14.14)$$

и в общем случае n_{n2} равен 2^{n-3} . Конечно, без более глубокого анализа мы не можем получить оценку для $\|N^{-1}F\|_2$, которая не содержала бы множителя 2^n . Можно сопоставить это с ситуацией, в которой N была бы унитарной. Тогда мы сразу имеем

$$\|N^{-1}F\|_2 = \|F\|_2.$$

15. Таким образом, процесс имеет два источника опасности. Во-первых, элементы H могут быть значительно больше элементов A , и тогда не выполняются предположения, на которых основана наша оценка для F . Если наибольший элемент H будет порядка 2^k , то оценка для F будет содержать дополнительный множитель 2^k . Во-вторых, даже если мы используем выбор ведущего элемента, норма N^{-1} может быть весьма велика. Ситуация похожа на ту, какую мы имели в методе Гаусса (или при разложении матрицы в произведение треугольных) с выбором ведущего

элемента по столбцу, хотя теперь мы имеем дополнительную опасность от нормы N^{-1} . Анализ скорости роста ведущих элементов до сих пор не проведен.

На практике наблюдался лишь небольшой рост ведущих элементов, и матрицы N^{-1} имели нормы порядка единицы. Следовательно, $\|N^{-1}F\|_2$ была обычно меньше, чем $\frac{1}{2}n2^{-t}$, что является оценкой, которую мы нашли для $\|F\|_2$. Снова ситуация сильно напоминает ситуацию, которую мы описывали в связи с методом Гаусса.

По существу преобразования, использующие устойчивые элементарные матрицы, обычно оказываются даже более точными, чем метод Хаусхолдера. Однако остается потенциальная опасность роста ведущих элементов и нормы N^{-1} , и из предосторожности следует иногда предпочесть метод Хаусхолдера. Заметим, что сейчас имеется несколько больше оправданий для осторожной точки зрения, чем в случае приведения к треугольному виду. Если имеется вычислительная машина, на которой скалярные произведения не могут удобно накапливаться, то использование устойчивых элементарных преобразований значительно менее привлекательно.

Дальнейший анализ ошибок

16. Мы дали анализ ошибок, соответствующий использованию арифметики с фиксированной запятой и накоплением скалярных произведений. Если использовать плавающую запятую с накоплением, пренебрегая членами порядка 2^{-2t} , то, например, для случая $n = 5$ получим для F оценку вида

$$|F| < 2^{-t} \begin{bmatrix} 0 & |h_{12}| & |h_{13}| & |h_{14}| & |h_{15}| \\ |h_{21}| & |h_{22}| & |h_{23}| & |h_{24}| & |h_{25}| \\ |h_{31}| & |h_{32}| & |h_{33}| & |h_{34}| & |h_{35}| \\ |h_{41}| & |h_{42}| & |h_{43}| & |h_{44}| & |h_{45}| \\ |h_{51}| & |h_{52}| & |h_{53}| & |h_{54}| & |h_{55}| \end{bmatrix}. \quad (16.1)$$

Как обычно, мы несколько больше выигрываем от малости элементов преобразованной матрицы по сравнению с вычислениями с фиксированной запятой, помимо обычного преимущества плавающей запятой — отсутствия нормировок в случае роста ведущих элементов.

Анализ методов §§ 8 и 9 имеет много общего с анализом прямого приведения к форме Хессенберга. Для методов § 9, например для модификации I с фиксированной запятой и накоплением скалярных произведений, мы имеем

$$(I_{r+1, (r+1)'} A_{r-1} I_{r+1, (r+1)'} N_{r+1}^{-1} - N_{r+1}^{-1} A_r \equiv F_r. \quad (16.2)$$

Из (9.1) и из уравнений, определяющих $n_i, r+1$, находим, например, в случае $n = 6, r = 2$, что

$$|F_r| \leq 2^{-t-1} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \quad (16.3)$$

Снова предполагается, что все элементы A_{r-1} и A_r ограничены по модулю единицей.

Уравнение (16.2) может быть записано в виде

$$N_{r+1} (I_{r+1, (r+1)'} A_{r-1} I_{r+1, (r+1)'}) N_{r+1}^{-1} = A_r + N_{r+1} F_r = A_r + G_r, \quad (16.4)$$

так что

$$|G_r| \leq 2^{-t-1} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 1 \end{bmatrix}. \quad (16.5)$$

Аналогично для модификации II можно показать, что

$$N_{r+1} (I_{r+1, (r+1)'} A_{r-1} I_{r+1, (r+1)'}) + K_r) N_{r+1}^{-1} \equiv A_r, \quad (16.6)$$

и для K_r можно получить ту же оценку, что и для G_r . Уравнение (16.6) дает возмущение A_{r-1} , эквивалентное ошибкам на r -м шаге.

К сожалению, для любой модификации окончательная оценка возмущения A_0 , эквивалентного ошибкам округления в полном процессе, содержит множитель 2^n и, как можно было ожидать, несколько хуже, чем оценка, полученная для прямого приведения.

17. Полезно проследить историю отдельного собственного значения λ матрицы A_0 в процессе перехода от A_0 к A_{n-2} . Вообще говоря, число обусловленности (гл. 2, § 34) этого собственного значения будет меняться, и если обозначить соответствующую величину через $1/|s^{(i)}|$ ($i = 0, 1, \dots, n-2$), то грубые верхние границы для окончательных ошибок в λ будут соответственно

$$\sum_{r=1}^{n-2} \|G_r\|_2 / |s^{(r)}| \quad \text{и} \quad \sum_{r=1}^{n-2} \|K_r\|_2 / |s^{(r-1)}| \quad (17.1)$$

при использовании модификаций I и II. Если мы обозначим

$$\alpha = \max |s^{(0)}/s^{(r)}|, \quad (17.2)$$

то окончательные ошибки в λ будут не больше, чем ошибки от возмущений A_0 , ограниченных соответственно величинами

$$\alpha \sum_{r=1}^{n-2} \|G_r\|_2 \quad \text{и} \quad \alpha \sum_{r=1}^{n-2} \|K_r\|_2, \quad (17.3)$$

и, следовательно, если использовать при оценках евклидову норму и неравенство (16.5), величиной

$$\alpha 2^{-t-1} \sum_{r=1}^{n-2} [r(r+4) + n]^{1/2}. \quad (17.4)$$

Эта сумма асимптотически стремится к $\frac{1}{2}n^2$.

Если числа обусловленности каждого собственного значения не меняются в слишком широких пределах, α будет порядка единицы, и полученный результат весьма удовлетворителен. Если y и x — нормированные левый и правый собственные векторы, соответствующие λ , то собственными

векторами A_1 будут $N_2^T y$ и $N_2^{-1} x$, но эти векторы уже не нормированы. Поэтому мы имеем

$$\left| \frac{s_0}{s_1} \right| = \left| \frac{(y^T x) \|N_2^T y\|_2 \|N_2^{-1} x\|_2}{y^T N_2 N_2^{-1} x} \right| = \|N_2^T y\|_2 \|N_2^{-1} x\|_2. \quad (17.5)$$

Следовательно, отношение s_0 к s_1 зависит от длины векторов $N_2^T y$ и $N_2^{-1} x$, компоненты которых равны

$$\left. \begin{aligned} (y_1, y_2 + n_{32}y_3 + n_{42}y_4 + \dots + n_{n2}y_n, y_3, \dots, y_n), \\ (x_1, x_2, x_3 - n_{32}x_2, x_4 - n_{42}x_2, \dots, x_n - n_{n2}x_2) \end{aligned} \right\} \quad (17.6)$$

соответственно. Длины преобразованных собственных векторов могут, следовательно, постоянно расти, даже если в результате выбора ведущих элементов мы имеем $|n_{ij}| \leq 1$.

Мы просчитали эти изменения для случайно выбранных матриц и нашли, что изменения чисел обусловленности были удивительно малы. Однако из-за большого объема работы, количество примеров было невелико.

Подчеркиваем, что если выбор ведущих элементов не используется и мы имеем $n_{i, r+1}$ порядка 2^k , то, вообще говоря, имеются элементы A_r , которые больше элементов A_{r-1} в 2^{2k} раз. Эквивалентное возмущение исходной матрицы также больше во столько же раз. Неприятности от неправильного выбора ведущих элементов поэтому могут быть даже более серьезными, чем в методе Гаусса. Кроме того, наше обсуждение изменений чисел обусловленности отдельного собственного значения показывает, что если $n_{i, r+1}$ велик, то в результате r -го преобразования возможно резкое ухудшение числа обусловленности.

Плохая определимость матрицы Хессенберга

18. Если на какой-либо стадии приведения, описанного в § 12, все элементы σ_i оказались малыми в результате вычитания, то ведущий элемент, который равен $\max |\sigma_i|$, будет мал. Этот ведущий элемент должен быть округлен перед тем, как на него будут делиться другие σ_i для получения множителей $n_{i, r+1}$, и, следовательно, все эти множители плохо определены. В этом случае окончательно вычисленная матрица Хессенберга сильно зависит от точности вычислений. Важно ясно представлять себе, что эта плохая определимость H никоим образом не связана с плохой определемостью ее собственных значений. Эта ситуация полностью аналогична той, которая обсуждалась в гл. 5, § 28 в связи с плохой определемостью окончательной трехдиагональной матрицы при симметричном приведении Гивенса. Полезно рассмотреть, что случится, если при точных вычислениях все σ_i равны нулю. В этом случае соответствующие $n_{i, r+1}$ не определены, и мы можем их взять произвольно. Соответственно этой полной произвольности в случае точных вычислений мы имеем плохую определимость, когда σ_i малы на практике. Эти рассуждения будут полезны далее.

Приведение к форме Хессенберга матрицами типа M'_{ji}

19. В гл. 4, § 48 был описан метод приведения матриц к треугольному виду при помощи умножения слева на матрицы типа M'_{ji} . Существует соответствующий метод преобразования подобия к форме Хессенберга. Как и в случае приведения к треугольному виду, он является близким

аналогом метода, основанного на плоских вращениях; $(n - r - 1)$ нулей в r -м столбце получаются один за другим при помощи преобразования с $(n - r - 1)$ устойчивыми матрицами $M'_{j, r+1}$ вместо одной устойчивой матрицы N'_{r+1} .

В случае приведения к треугольному виду этот метод можно было рекомендовать, так как он позволял найти все ведущие главные миноры исходной матрицы. При подобном преобразовании таких преимуществ нет, и так как число умножений здесь такое же, как и при прямом приведении § 11, мы не будем обсуждать этот метод детально.

Метод Крылова

20. Опишем теперь алгоритм (обычно используемый для вычисления коэффициентов характеристического полинома), который естественным образом приведет нас ко второму классу методов, тесно связанных с уже описанными. Этот метод принадлежит Крылову (1931).

Если b_0 — вектор высоты n , то векторы b_r :

$$b_r = A^r b_0 \quad (20.1)$$

удовлетворяют уравнению

$$b_n + p_{n-1}b_{n-1} + \dots + p_0b_0 = 0, \quad (20.2)$$

где $\lambda^n + p_{n-1}\lambda^{n-1} + \dots + p_0$ — характеристический полином A (гл. 1, §§ 34—39). Уравнение (20.2) может быть записано в виде

$$Bp = -b_n, \quad B = (b_0 | b_1 | \dots | b_{n-1}). \quad (20.3)$$

Это дает неособенную систему линейных уравнений для определения p , и, следовательно, для вычисления коэффициентов характеристического полинома A . Если b_0 — вектор высоты, меньшей чем n , то минимальный полином b_0 по отношению к A будет меньшей степени чем n . Тогда

$$b_m + c_{m-1}b_{m-1} + \dots + c_0b_0 = 0 \quad (m < n), \quad (20.4)$$

и из уравнения (20.4) можно определить коэффициенты минимального аннулирующего b_0 полинома $\lambda^m + c_{m-1}\lambda^{m-1} + \dots + c_0$; этот полином будет множителем характеристического полинома. Если A — неполная, то высота каждого вектора будет меньше n .

На практике мы заранее не будем знать, является ли b_0 вектором высоты, меньшей n . Желательно было бы использовать способы, которые давали бы эту информацию автоматически: Процесс, который мы сейчас будем описывать, основан на следующей модификации метода Гаусса.

Метод Гаусса с исключением по столбцам

21. Обозначим матрицу, которая приводится к верхнему треугольному виду, через X , а ее столбцы через x_i . Процесс состоит из n основных шагов, по одному для каждого отдельного столбца, причем столбцы $x_r, x_{r+1}, \dots, x_n$ не меняются при первых $(r - 1)$ шагах, в конце

которых матрица X , например, для случая $n = 5$, $r = 3$, имеет вид

$$\begin{array}{ccccc|c} u_{11} & u_{12} & x_{13} & x_{14} & x_{15} & \sigma_1 \\ n_{21} & u_{22} & x_{23} & x_{24} & x_{25} & \sigma_2 \\ n_{31} & n_{32} & x_{33} & x_{34} & x_{35} & \sigma_3 \\ n_{41} & n_{42} & x_{43} & x_{44} & x_{45} & \sigma_4 \\ n_{51} & n_{52} & x_{53} & x_{54} & x_{55} & \sigma_5 \\ 1' & 2' & & & & \end{array} \quad (21.1)$$

Основной r -й шаг состоит в следующем.

(i) Для всех значений i от 1 до n умножаем x_{ir} на 1 и записываем произведение в ячейке двойной точности σ_i . Для всех значений i от 1 до $r - 1$ выполняем шаги (ii) и (iii):

(ii) Переставляем σ_i и σ_r , округляем новое σ_i до одинарной точности для получения u_{ir} и записываем на месте x_{ir} .

(iii) Для всех значений j от $i + 1$ до n заменяем σ_j на $\sigma_j - n_{ji}u_{ir}$.

(iv) Выбираем $\sigma_{r'}$ так, что $|\sigma_{r'}| = \max_{i=r}^n |\sigma_i|$. Если таких максимумов более одного, выбираем первый. Записываем r' внизу r -го столбца.

(v) Переставляем σ_r и $\sigma_{r'}$, округляем новое σ_r до одинарной точности для получения u_{rr} и записываем на месте x_{rr} .

(vi) Для всех значений i от $r + 1$ до n вычисляем $n_{ir} = \sigma_i/u_{rr}$ и записываем на месте x_{ir} .

Подчеркнем, что этот процесс совпадает с треугольным разложением, использующим частичный выбор главного элемента и накопление скалярных произведений (гл. 4, § 39); даже ошибки округления те же самые. Однако если r -й столбец X линейно зависит от предыдущих, это проявится в течение r -го основного шага модифицированного процесса, через появление нулей σ_i ($i = r, \dots, n$) на вспомогательном шаге (iv). В этом случае не нужны никакие вычисления с последующими столбцами.

Описанный процесс можно применить к матрице B , образованной векторами $b_0, b_1, \dots, b_{n-1}$. Если при работе с b_m мы заметим, что b_m линейно зависим с $b_0, b_1, \dots, b_{m-1}$, то мы не будем вычислять остальные векторы. Тогда

$$U_m = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ & u_{22} & \dots & u_{2m} \\ & & \ddots & \\ & & & u_{mm} \end{bmatrix}, \quad u_m = \begin{bmatrix} u_{1, m+1} \\ u_{2, m+1} \\ \vdots \\ u_{m, m+1} \end{bmatrix}, \quad (21.2)$$

так что решение уравнений

$$U_m c = -u_m \quad (21.3)$$

дает коэффициенты $c_0, c_1, \dots, c_{m-1}$ минимального полинома вектора b_0 по отношению к A .

Практические трудности

22. Нетрудно показать, что если A полна, то почти все векторы высоты n и поэтому приводят к полному характеристическому полиному.

На практике различие между линейной зависимостью и независимостью затушевывается в результате ошибок округления. Даже если высота b_0 меньше n , вряд ли соответствующие σ_i точно исчезнут. Более

того, если матрица B плохо обусловлена, то могут возникнуть малые σ_i , и тогда трудно установить настоящую линейную зависимость. К сожалению, имеется ряд факторов, которые делают B плохо обусловленной. Мы сейчас их обсудим.

Для того чтобы разобраться в этих факторах, предположим, что A это $\text{diag}(\lambda_i)$ и что

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|. \quad (22.1)$$

Если начальный вектор b_0 имел компоненты $\beta_1, \beta_2, \dots, \beta_n$, и если какая-либо β_i равна нулю, то высота b_0 меньше n , и, следовательно, B — особенная. Поэтому мы можем ожидать, что малость β_i ведет к плохой обусловленности B , и наилучшим выбором будет $\beta_i = 1$ ($i = 1, \dots, n$). Матрица B тогда имеет вид

$$B = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \dots & \lambda_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \lambda_n & \lambda_n^2 & \dots & \lambda_n^{n-1} \end{bmatrix}, \quad (22.2)$$

и уравнение, определяющее коэффициенты характеристического полинома, будет

$$Bp = -b_n, \quad b_n^T = (\lambda_1^n, \lambda_2^n, \dots, \lambda_n^n). \quad (22.3)$$

Из явного выражения видно, что B — особенная, если $\lambda_i = \lambda_j$ (как мы уже знаем), и плохо обусловлена, если какие-либо два собственных значения близки. Мы можем уравновесить B , перейдя к C вида

$$C = \begin{bmatrix} 1 & \mu_1 & \mu_1^2 & \dots & \mu_1^{n-1} \\ 1 & \mu_2 & \mu_2^2 & \dots & \mu_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \mu_n & \mu_n^2 & \dots & \mu_n^{n-1} \end{bmatrix}, \quad \text{где } \mu_i = \lambda_i / \lambda_1, \quad (22.4)$$

и число обусловленности C оценивается наибольшим элементом ее обратной матрицы. Матрица C это матрица Вандермонда, и можно показать, что если Z ее обратная, то

$$z_{sr} = p_{rs} / \prod_{i \neq r} (\mu_r - \mu_i), \quad (22.5)$$

где p_{rs} определяются из соотношения

$$\prod_{i \neq r} (y - \mu_i) = \sum_{s=0}^{n-1} p_{r, s+1} y^s. \quad (22.6)$$

Обусловленность C для некоторых стандартных расположений собственных значений

23. Сейчас мы оценим числа обусловленности C для нескольких стандартных расположений собственных значений матрицы порядка 20.

(i) Линейное расположение $\lambda_r = 21 - r$. Имеем

$$|1 / \prod_{i \neq r} (\mu_r - \mu_i)| = 20^{19} / (r-1)! (20-r)!, \quad (23.1)$$

и наибольшее значение, приблизительно равное $4 \cdot 10^{12}$, достигается при $r = 10$ или 11 . Так как некоторые коэффициенты $p_{10,s}$ порядка 10^3 , мы видим, что некоторые элементы Z превосходят 10^{15} .

(ii) Линейное расположение $\lambda_r = 11 - r$. Теперь мы имеем

$$|1/\prod_{i \neq r} (\mu_r - \mu_i)| = 10^{19}/(r-1)!(20-r)! \quad (23.2)$$

Это на множитель 2^{-19} меньше значения, полученного для расположения (i), и соответствующие величины $|p_{rs}|$ также меньше. Число обусловленности меньше по крайней мере в 10^6 раз.

(iii) Геометрическое расположение $\lambda_r = 2^{-r}$. Для $r = 20$ имеем

$$\begin{aligned} |1/\prod_{i \neq r} (\mu_r - \mu_i)| &= 2^{-19}/(2^{-1} - 20^{-20})(2^{-2} - 2^{-20}) \dots (2^{-19} - 2^{-20}) > \\ &> 2^{-19}/2^{-1}2^{-2} \dots 2^{-19} = 2^{171}. \end{aligned} \quad (23.3)$$

Теперь $p_{20,1}$ почти равен 1, так что число обусловленности C порядка 2^{171} .

(iv) Расположение на окружности $\lambda_r = \exp(r\pi i/10)$. Для этого расположения имеем

$$|1/\prod_{i \neq r} (\mu_r - \mu_i)| = 1/20 \quad (\text{для всех } r). \quad (23.4)$$

Кроме того, все p_{rs} по модулю равны единице. Поэтому матрица очень хорошо обусловлена, и мы можем ожидать, что коэффициенты характеристического уравнения определятся с максимальной возможной точностью. (Заметим, что λ_i комплексные, и если предполагается, что A диагональна, необходимо применение комплексной арифметики.)

Если A не диагональная, но

$$A = H^{-1} \text{diag}(\lambda_i) H, \quad (23.5)$$

то

$$b_r = A^r b_0 = H^{-1} \text{diag}(\lambda_i^r) (H b_0). \quad (23.6)$$

Последовательность, соответствующая матрице A и начальному вектору $b_0 = H^{-1}c_0$, совпадает с последовательностью, соответствующей $\text{diag}(\lambda_i)$ и начальному вектору c_0 , с точностью до множителя H^{-1} . Поэтому, естественно, мы не можем ожидать, что матрица, соответствующая A , будет сколько-нибудь лучше обусловлена, чем матрица, соответствующая $\text{diag}(\lambda_i)$. Следовательно, возвращаясь к значению использования e в качестве начального вектора, можно ограничиться снова диагональными матрицами.

Очевидно, что если величины собственных значений сильно различаются, то последние векторы последовательности имеют очень малые последние компоненты. Следующий пример показывает, что эта трудность фундаментальная, и что любой поиск такого b_0 , чтобы B была хорошо обусловлена, обречен на неудачу.

Для расположения (i) можно рассмотреть эффект от использования в качестве начального вектора

$$b_0^T = (20^{-10}, 19^{-10}, \dots, 1), \quad (23.7)$$

в котором компоненты, соответствующие малым собственным значениям, хорошо представлены. Тогда имеем

$$b_{10}^T = (1, 1, 1, \dots, i) \quad (23.8)$$

и

$$b_{20}^T = 20^{10} [1, (19/20)^{10}, (18/20)^{10}, \dots, (1/20)^{10}], \quad (23.9)$$

где хорошо представлены компоненты, соответствующие большим собственным значениям. Соответствующая уравновешенная матрица C имеет вид

$$\begin{bmatrix} 20^{-10} & 20^{-9} & \dots & 1 & 1 & \dots & 1 \\ 19^{-10} & 19^{-9} & \dots & 1 & 19/20 & \dots & (19/20)^9 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 2^{-10} & 2^{-9} & \dots & 1 & 2/20 & \dots & (2/20)^9 \\ 1 & 1 & \dots & 1 & 1/20 & \dots & (1/20)^9 \end{bmatrix} = C'. \quad (23.10)$$

Очевидно, что если D_1 и D_2 — диагональные матрицы вида

$$D_1 = [1, (19/20)^{10}, (18/20)^{10}, \dots, (1/20)^{10}], \quad (23.11)$$

$$D_2 = (20^{10}, 20^9, \dots, 20^1, 1, 1, \dots, 1), \quad (23.12)$$

то $D_1 C' D_2$ совпадает с матрицей, которая бы соответствовала $b_0 = e$. Следовательно, новая обратная Y выражается через старую обратную при помощи соотношения

$$Y = D_2 Z D_1, \quad (23.13)$$

и Y все равно очень плохо обусловлена.

Начальный вектор высоты, меньшей чем n

24. Полезно рассмотреть, что получится, если в качестве начального вектора взять вектор высоты m , меньшей чем n . Рассмотрим сначала этот вопрос с теоретической точки зрения. Имеем

$$\sum_{i=0}^{m-1} c_i b_i + b_m = 0 \quad (m < n), \quad (24.1)$$

и, следовательно, умножая слева на A^s ,

$$\sum_{i=0}^{m-1} c_i b_{i+s} + b_{m+s} = 0 \quad (s = 0, \dots, n-m). \quad (24.2)$$

Это показывает, что любые $m+1$ последовательных b_r удовлетворяют тому же самому линейному соотношению. Очевидно, что уравнения

$$Bp = -b_n \quad (24.3)$$

совместны, так как $\text{rank}(B) = \text{rank}(B; b_n) = m$ из соотношений (24.2). Следовательно, мы можем выбрать $p_m, p_{m+1}, \dots, p_{n-1}$ произвольно, и затем определить $p_0, \dots, p_{m-1}$. Общая форма решения может быть получена из уравнений (24.2). Действительно, если мы возьмем произвольные значения q_s ($s = 0, \dots, n-m-1$) и $q_{n-m} = 1$, умножим соответствующие уравнения (24.2) на q_s и сложим, то

$$\sum_{i=0}^{n-1} p_i b_i + b_n = 0, \quad (24.4)$$

где p_i — коэффициенты при x^i в разложении

$$(c_0 + c_1x + \dots + c_{m-1}x^{m-1} + x^m)(q_0 + q_1x + \dots + q_{n-m-1}x^{n-m-1} + x^{n-m}).$$

Так как мы имеем $n - m$ произвольных постоянных, p_i представляют общее решение (24.3). Соответствующий полином

$$p_0 + p_1\lambda + \dots + p_{n-1}\lambda^{n-1} + \lambda^n,$$

следовательно, равен

$$(c_0 + c_1\lambda + \dots + c_{m-1}\lambda^{m-1} + \lambda^m)(q_0 + q_1\lambda + \dots + q_{n-m-1}\lambda^{n-m-1} + \lambda^{n-m}) \quad (24.5)$$

и он имеет в качестве сомножителя минимальный полином b_0 при произвольных остальных сомножителях.

На практике, если мы не различим наступление линейной зависимости и будем продолжать процесс до получения b_n , при выполнении приведения к треугольному виду по способу § 21 треугольник порядка $(n - m)$ в нижнем правом углу u_n будет состоять полностью из «малых» элементов, а не из нулей; последние $(n - m)$ элементов u_n также будут малыми.

Следовательно, мы получим более или менее случайные значения $p_{n-1}, p_{n-2}, \dots, p_m$. Однако величины $p_0, p_1, \dots, p_{m-1}$ будут определены хорошо из первых m уравнений треугольной системы. Следовательно, вычисленный полином будет иметь множитель

$$c_0 + c_1\lambda + \dots + c_{m-1}\lambda^{m-1} + \lambda^m,$$

а другой сомножитель степени $(n - m)$ будет весьма произвольным.

Практическое применение

25. Результаты, описанные в нескольких предыдущих параграфах, были подтверждены на практике, причем использовалась программа, написанная для DEUCE. С целью компенсации ожидавшейся плохой обусловленности матрицы B в процессе вычислений использовалась арифметика с двойной точностью и фиксированной запятой.

Вычислялись две последовательности $b_0, b_1, \dots, b_n$ и $c_0, c_1, \dots, c_n$ из соотношений

$$b_0 = c_0, \quad b_{r+1} = Ac_r, \quad c_{r+1} = k_{r+1}b_{r+1}, \quad (25.1)$$

где k_{r+1} выбиралось так, чтобы $\|c_{r+1}\|_\infty = 1$. Исходная матрица A имела элементы только одинарной точности и при образовании Ac_r скалярные произведения накапливались точно с округлением после суммирования; это требовало лишь умножения величин с двойной точностью (62 двоичных разряда) на величины с одинарной точностью. Так как исходная A используется все время, мы можем полностью получать все преимущества от нулевых элементов, которые она может содержать. Уравнения

$$(c_0 | c_1 | \dots | c_{n-1})q = -c_n \quad (25.2)$$

затем решались с использованием обычной арифметики с двойной точностью и фиксированной запятой. и окончательно коэффициенты

характеристического уравнения вычислялись из соотношений

$$p_r k_{r+1} k_{r+2} \cdot \dots \cdot k_n = q_r \quad (25.3)$$

с использованием арифметики с двойной точностью и плавающей запятой.

Предпринимались попытки выявления, как в § 21, линейной зависимости на самой ранней стадии; при этом заменялись нулями все элементы, меньшие по модулю некоторого предписанного числа. (В этом случае удобно нормировать b_r в c_r , даже если все время используется *плавающая запятая* с двойной точностью.) Для того чтобы вызвать наступление линейной зависимости, я так же пытался несколько раз предварительно умножать исходный вектор на A , чтобы начальный вектор имел малые составляющие по собственным векторам, соответствующим меньшим собственным значениям.

В целом мой опыт с этими методами был не очень удачным, и я отказался от их использования в пользу вычисления во всех случаях полной системы векторов. Результаты, полученные на практике, следующие:

(i) Для диагональной матрицы $\text{diag}(\lambda_i)$ с $\lambda_r = 21 - r$ при начальном векторе e характеристическое уравнение имело 13 правильных двоичных знаков в наиболее точных коэффициентах и 10 в остальных. Мы могли ожидать такую неточность, так как число обусловленности C порядка 10^{15} . Используя начальный вектор (23.7), мы получили наиболее точные коэффициенты с 15 правильными двоичными знаками, а остальные с 11. Все выведенные уравнения совершенно бесполезны для определения собственных значений (см. гл. 7, § 6.)

(ii) Для диагональной матрицы порядка 21 с $\lambda_r = 11 - r$ при начальном векторе e коэффициенты вычисленного характеристического полинома имели по крайней мере 30 правильных двоичных знаков, причем некоторые коэффициенты и значительно больше. Это эквивалентно определению собственных значений приблизительно с 20 двоичными знаками.

(iii) Для расположения $\lambda_r = 2^{-r}$ вычисленные характеристические уравнения, соответствующие нескольким различным начальным векторам, были никак не связаны с правильным уравнением.

(iv) Для опыта с расположением $\lambda_r = \exp(r\pi i/10)$ без использования комплексной арифметики, мы брали трехдиагональную матрицу, имеющую 10 блоков вида

$$\begin{bmatrix} \cos \theta_r & \sin \theta_r \\ -\sin & \cos \theta_r \end{bmatrix}, \quad \theta_r = r\pi/10. \quad (25.4)$$

При начальном векторе e это дало характеристическое уравнение, правильное с точностью до 60 двоичных знаков; собственные значения были правильными с той же точностью.

Для простых расположений с несколькими совпадающими или очень близкими собственными значениями обычно получались несколько правильных собственных значений, а некоторые были довольно произвольными, что подтверждает анализ § 24.

Обобщенные процессы Хессенберга

26. Наше обсуждение метода Крылова показало, что он дает хорошие результаты только при благоприятных расположениях собственных значений. Однако надо заметить, что такие благоприятные расположения совсем не необычны для матриц, возникающих в связи с затухающими

механическими и электрическими колебаниями, когда собственные значения являются, вообще говоря, комплексно сопряженными парами.

С целью расширения круга удовлетворительных приложений были найдены методы, которые «стабилизировали» последовательность векторов Крылова. Они могут быть объединены под названием «обобщенных методов Хессенберга». Основа всех таких методов заключается в следующем.

Пусть дана система n линейно независимых векторов x_i , которые являются столбцами матрицы X . Начиная с произвольного вектора b_1 , мы образуем систему модифицированных векторов Крылова $b_2, b_3, \dots, b_{n+1}$, определенных соотношениями

$$k_{r+1}b_{r+1} = Ab_r - \sum_{i=1}^r h_{ir}b_i \quad (26.1)$$

(Нам удобнее называть теперь начальный вектор b_1 , а не b_0 , так как мы уже не занимаемся прямым определением характеристического уравнения.) Здесь k_{r+1} — скаляры, введенные для численного удобства, являющиеся обычно нормирующими множителями. Числа h_{ir} определены так, чтобы b_{r+1} был ортогонален $x_1, \dots, x_r$. Вектор b_{n+1} должен быть равен нулю, если только не будет равен нулю более ранний b_r , так как он ортогонален к n линейно независимым векторам. Уравнения (26.1) могут быть записаны в виде одного матричного уравнения

$$A [b_1 | b_2 | \dots | b_n] = [b_1 | b_2 | \dots | b_n] \begin{bmatrix} h_{11} & h_{12} & h_{13} & \dots & h_{1n} \\ k_2 & h_{22} & h_{23} & \dots & h_{2n} \\ 0 & k_3 & h_{33} & \dots & h_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & k_n & h_{nn} \end{bmatrix}, \quad (26.2)$$

или

$$AB = BH, \quad (26.3)$$

где матрица H в верхней форме Хессенберга.

То, что уравнение (26.2) включает условие $b_{n+1} = 0$ можно увидеть, если приравнять n -е столбцы в обеих сторонах и сравнить с (26.1). Условие ортогональности может быть записано в виде

$$X^T B = L, \quad (26.4)$$

где L — нижняя треугольная матрица. Предполагая, что B — неособенная, мы можем переписать (26.3) в виде

$$B^{-1}AB = H^{\#} \quad (26.5)$$

Для удобства мы предположили, что векторы x_i даны заранее. Однако это нигде не требуется, и на самом деле мы можем вводить их вместе с b_i при помощи какого-либо целесообразного процесса.

Срыв обобщенного процесса Хессенберга

27. Только что описанный процесс может оборваться по двум различным причинам.

(i) Вектор b_{r+1} ($r < n$) может оказаться равным нулю. В этом случае мы можем заменить нулевой b_{r+1} произвольным вектором c_{r+1} , который ортогонален $x_1, \dots, x_r$. Мы можем сохранить существующие

обозначения, определив новое множество векторов, удовлетворяющих следующим соотношениям:

$$\left. \begin{aligned} c_1 &= b_1, & b_{r+1} &= Ac_r - \sum_{i=1}^r h_{ir} c_i, \\ k_{r+1} c_{r+1} &= b_{r+1} & (b_{r+1} &\neq 0), \end{aligned} \right\} \quad (27.1)$$

c_{r+1} — произвольный не нулевой вектор, ортогональный к $x_1, \dots, x_r$ ($b_{r+1} = 0$).

Если при b_{r+1} , равном нулю, мы возьмем $k_{r+1} = 0$, то соотношение $k_{r+1} c_{r+1} = b_{r+1}$ остается в силе.

В этих обозначениях имеем

$$AC = CH, \quad X^T C = L. \quad (27.2)$$

Каждому нулевому b_r соответствует нулевой k_r в H . Мы видим, что появление нулевых b_r вряд ли надо рассматривать как «срыв» процесса, а скорее как отправную точку для поисков методов продолжения процесса. В этом случае вычисление собственных значений H проще, так как, если $k_r = 0$, то собственные значения H являются собственными значениями двух меньших матриц. С этого места мы обсуждаем процесс в терминах c_i .

(ii) Элементы матрицы H выводятся из соотношения

$$h_{ir} = x_i^T A c_r / x_i^T c_i. \quad (27.3)$$

Если $x_i^T c_i = 0$, то мы не можем определить h_{ir} . Очевидно, что обращение в нуль $x_i^T c_i$ более серьезно, чем преждевременное обращение в нуль b_i . Если x_i не даны заранее, то можно попытаться выбрать x_i так, чтобы $x_i^T c_i$ не было нулем. Если x_i даны заранее, то обращение в нуль $x_i^T c_i$ будет, конечно, фатальным, и единственное возможное решение — это начать процесс сначала с другим вектором b_1 . Интересный вопрос состоит в том, как выбрать b_1 так, чтобы такого срыва не произошло.

Заметим, что из (27.2) мы имеем $l_{ii} = x_i^T c_i$. Если все $x_i^T c_i$ отличны от нуля, то L — неособенная, и, следовательно, C — неособенная.

Метод Хессенберга

28. Естественным выбором для X является единичная матрица. В этом случае $x_i = e_i$, и это есть метод Хессенберга (1941). Второе из уравнений (27.2) становится

$$L = X^T C = C, \quad (28.1)$$

так что C — сама нижняя треугольная, т. е. первые $(r - 1)$ компонент c_r равны нулю. Мы можем выбрать k_r так, чтобы r -я компонента c_r была равна единице, положив

$$c_r^T = (0, \dots, 0, 1, n_{r+1,r}, \dots, n_{nr}) \quad (28.2)$$

по причинам, которые далее станут очевидными. r -й шаг процесса состоит в следующем.

Вычисляем Ac_r , который, вообще говоря, не имеет нулевых компонент. Затем последовательно вычитаем кратное c_1 так, чтобы исчезла первая компонента, кратное c_2 так, чтобы исчезла вторая, ..., и, наконец, кратное c_r так, чтобы исчезла r -я компонента. Это даст вектор b_{r+1} . Затем выбираем k_{r+1} равным $(r + 1)$ -й компоненте b_{r+1} , так что c_{r+1} окажется в форме (28.2). Если b_{r+1} равен нулю, то мы можем взять в качестве c_{r+1}

произвольный вектор вида

$$(0, \dots, 0, 1, n_{r+2, r+1}, \dots, n_{n, r+1}), \quad (28.3)$$

так как он автоматически ортогонален $e_1, \dots, e_r$. Соответствующее значение k_{r+1} равно нулю. Это соответствует случаю (i) § 27.

Если b_{r+1} не равен нулю, но его $(r+1)$ -я компонента равна нулю, мы не можем нормировать b_{r+1} так, чтобы получить c_{r+1} в предписанной форме. При любом нормирующем множителе $e_{r+1}^T c_{r+1} = 0$, и следовательно, процесс сорвется на следующем шаге. Это соответствует случаю (ii) в § 27.

Практическая процедура

29. Мы можем избежать срыва, а также обеспечить численную устойчивость, если возьмем в качестве x_i столбцы I , но не обязательно в естественном порядке. Возьмем $x_r = e_{r'}$, где $1', 2', \dots, n'$ некоторая перестановка $1, 2, \dots, n$.

Способ, по которому определяется r' , вероятно, лучше всего объяснить, если описать r -й шаг. Пусть уже определены векторы $c_1, \dots, c_r$ и $1', \dots, r'$ и c_i имеет нули в позициях $1', \dots, (i-1)'$. r -й шаг состоит в следующем.

Вычисляем Ac_r и вычитаем кратные $c_1, c_2, \dots, c_r$ так, чтобы аннулировать элементы $1', 2', \dots, r'$ соответственно. В результате получим вектор b_{r+1} . Если наибольший элемент b_{r+1} находится в позиции $(r+1)'$, то берем этот элемент в качестве k_{r+1} , а x_{r+1} будет $e_{(r+1)'}$. Очевидно, что элементы $1', \dots, r'$ вектора c_{r+1} — нули и что все отличные от нуля элементы ограничены единицей. Если b_{r+1} равен нулю, то мы можем взять в качестве x_{r+1} любой столбец $e_{(i+1)'}$ матрицы I , ортогональный e_i ($i = 1, \dots, r$), а в качестве c_{r+1} — произвольный вектор с единичной $(r+1)$ -й компонентой, также ортогональный e_i ($i = 1, \dots, r$).

Очевидно, что e_i образуют столбцы матрицы перестановок P (гл. 1, § 40 (ii)). Уравнения (27.2) тогда станут

$$AC = CH, \quad P^T C = L, \quad (29.1)$$

где L — не только нижняя треугольная, а единичная нижняя треугольная. Следовательно, имеем

$$P^T A (PP^T) C = P^T CH \quad (29.2)$$

или

$$(P^T A P) (P^T C) = (P^T C) H, \quad (29.3)$$

т. е.

$$\tilde{A} L = L H, \quad (29.4)$$

где $\tilde{A}$ это подобное преобразование A с помощью матрицы перестановок, а L это C с переставленными строками.

Связь метода Хессенберга и предыдущих методов

30. Если векторы e_i используются в естественном порядке, то мы имеем

$$AC = CH, \quad (30.1)$$

и, если c_r даются уравнением (28.2), мы можем написать

$$AN = NH, \quad (30.2)$$

где N — единичная треугольная матрица. Если же мы возьмем в качестве начального вектора c_1 вектор e_i , то из § 11 мы видим, что матрицы N

и H совпадают с теми, которые получаются при прямом приведении к форме Хессенберга. Использование в качестве c_1 какого-либо другого вектора с единичной первой компонентой эквивалентно выполнению предварительного преобразования $N_1 A N_1^{-1}$.

Практическая процедура, описанная в § 29, в точности эквивалентна прямому приведению к форме Хессенберга с включением перестановок. Заметим, что r' в § 29 это не r' в § 12, который таков, что $I_{2,2} I_{3,3'}, \dots, \dots, I_{n-1,(n-1)'} = P$. В самом деле, легко так переделать описание § 29, чтобы выполнять на каждом этапе одинаковые перестановки строк и столбцов A и строк текущей матрицы из c_i , вместо того чтобы использовать e_i в некотором переставленном порядке. Эти два метода не только математически эквивалентны, но если применять всюду, где возможно, накопление скалярных произведений, то даже ошибки округления совпадают.

Поэтому нам не надо проводить отдельный анализ ошибок для метода Хессенберга. Однако мы опишем ранее полученные результаты применительно к этому методу. Для удобства описания предположим, что A и векторы действительно переставляются на каждом этапе.

(i) Использование перестановок существенно в том смысле, что в противном случае мы могли бы получить бесконечно большие элементы n_{ji} вектора c_i , и это может привести к большей потере точности.

(ii) Если на какой-либо стадии вектор b_i имеет малые элементы во всех позициях от i до n , то от этого не происходит никакой потери точности, несмотря на то, что n_{ij} вычисляются как отношения малых величин. Мы просто имеем плохую определенность окончательной матрицы H . Мы умышленно подчеркиваем это снова; уже говорилось, что в этом случае надо начинать с другого начального вектора в надежде, что такая ситуация не будет повторяться.

(iii) Если вычисленный вектор b_{r+1} равен нулю, то мы можем взять $c_{r+1} = e_{r+1}$. Это соответствует тому факту, что если $a_{ir}^{(r-1)} = 0$ ($i = r+1, \dots, n$), то матрица N_{r+1} из § 8 может быть выбрана единичной, хотя, конечно, можно использовать любые N_{r+1} .

Метод Арнольди

31. Другим естественным выбором системы векторов x_i в обобщенном процессе Хессенберга является система самих c_i . Это ведет к методу Арнольди (1951). Второе уравнение (27.2) тогда станет

$$C^T C = D; \quad (31.1)$$

где D — диагональная, так как она одновременно нижняя треугольная и симметричная. Мы можем выбрать нормирующий множитель так, чтобы $C^T C = I$, и процесс тогда дает ортогональную матрицу. (Преимущества этой нормировки незначительны по сравнению с такой нормировкой, когда наибольший элемент каждого $|c_i|$ лежит между $1/2$ и 1 . Последняя нормировка более экономична. Однако мы увидим, что нормировка $C^T C = I$ обнаруживает интересную связь с предыдущими методами.) Предположим пока, что C не симметричная, так как симметричный случай обсуждается в другом контексте в § 41.

Если мы возьмем c_1 равным e_1 , то имеем

$$A C = C H \quad (31.2)$$

с ортогональной матрицей C , первый столбец которой равен e_1 . Следовательно, из § 7 мы видим, что C и H по существу суть матрицы, получающиеся при помощи преобразований Гивенса и Хаусхолдера.

Практические рассмотрения

32. Несмотря на то, что процесс Арнольди с нормировкой $C^TC = I$ так тесно связан с методами Гивенса и Хаусхолдера, вычислительный процесс имеет с ними очень мало общего. Рассмотрим более детально практический процесс. Нам больше не надо предполагать, что c_1 равен e_1 . Будем брать в качестве c_1 произвольный нормированный вектор. Рассмотрим вектор b_{r+1} вида

$$b_{r+1} = Ac_r - h_{1r}c_1 - h_{2r}c_2 - \dots - h_{rr}c_r, \quad (32.1)$$

где h_{ir} выбраны так, чтобы b_{r+1} был ортогонален к c_i ($i = 1, \dots, r$). Может случиться, что при вычислении b_{r+1} происходит взаимное уничтожение, так что все компоненты b_{r+1} малы по сравнению с $\|Ac_r\|_2$. В этом случае b_{r+1} не будет даже приблизительно ортогонален c_i . (Сравните с аналогичным явлением, обсуждавшимся в гл. 4, § 55 в связи с ортогонализацией Шмидта.)

Существенно, чтобы c_i оставались строго ортогональными с рабочей точностью. В противном случае нет гарантий, что c_{n+1} есть нуль с рабочей точностью. *Возникающую трудность нельзя отнести на счет накопления ошибок округления.* Это может случиться на первом шаге вычислений с матрицей второго порядка, что и иллюстрируется в табл. 3.

Таблица 3

A

$$\begin{bmatrix} 0,41323 & 0,61452 \\ 0,54651 & 0,21374 \end{bmatrix}$$

Процесс Арнольди без переортогонализации

c_1	Ac_1	b_2	
0,78289	0,70584	0,00004	
0,62216	0,56084	-0,00006	
$h_{11} = 0,90153$	$k_2 = 10^{-4}$	(0,72111)	
c_2	Ac_2	$Ac_2 - h_{12}c_1$	$b_3 = Ac_2 - h_{12}c_1 - h_{22}c_2$
0,55470	-0,28209	-0,17023	-0,01899
-0,83205	0,12531	0,21420	-0,01266
$h_{12} = -0,14288$		$h_{22} = -0,27265$	

H

$$\begin{bmatrix} 0,90153 & -0,14288 \\ 10^{-4} (0,72111) & -0,27265 \end{bmatrix}$$

Процесс Арнольди с переортогонализацией

c_1	Ac_1	b_2	$b'_2 = b_2 - e_{11}c_1$
0,78289	0,70584	0,00004	10^{-4} (0,44708)
0,62216	0,56084	-0,00006	10^{-4} (-0,56258)
	$h_{11} = 0,90153$	$k_2 = 10^{-4}$	(0,71859)
c_2	Ac_2	$Ac_2 - h_{12}c_1$	$b_3 = Ac_2 - h_{12}c_1 - h_{22}c_2$
0,62216	-0,22401	-0,17082	0,00000
-0,78289	0,17268	0,21495	0,00000
$h_{12} = -0,06794$		$h_{22} = -0,27456$	

H

$$\begin{bmatrix} 0,90153 & -0,06794 \\ 10^{-4} (0,71859) & -0,27456 \end{bmatrix}$$

При вычислении

$$b_2 = A c_1 - h_{11} c_1$$

происходит почти полное взаимное уничтожение. Следовательно, b_2 никоим образом не ортогонален c_1 . Если мы, пренебрегая этим, продолжим процесс, нормируя b_2 , то b_3 не будет почти нулем, как должно было бы быть, если бы c_1 и c_2 были почти точно ортогональны. Матрица H никоим образом не подобна A . (Вычисления были проведены с использованием блочно-плавающей арифметики, но это не имеет никакого значения. Те же самые трудности возникли бы при использовании любой арифметики.) След равен 0,62888 вместо 0,62697, и $\|b_3\|_2$ порядка $2 \cdot 10^{-2}$ вместо 10^{-5} .

Предположим теперь, что мы переортогонализируем вычисленный вектор b_2 по отношению к c_1 . Можно написать

$$b'_2 = b_2 - \varepsilon_{11} c_1, \quad (32.2)$$

и из ортогональности имеем

$$\varepsilon_{11} = c_1^T b_2. \quad (32.3)$$

Так как b_2 уже один раз ортогонализовался по отношению к c_1 , мы можем быть уверены, что ε_{11} порядка 10^{-5} . Следовательно,

$$b'_2 = b_2 + O(10^{-5}). \quad (32.4)$$

Исходная ортогонализация дает вектор b_2 , удовлетворяющий соотношению

$$b_2 = A c_1 - h_{11} c_1 + O(10^{-5}) \quad (32.5)$$

и, следовательно,

$$b'_2 = A c_1 - h_{11} c_1 + O(10^{-5}). \quad (32.6)$$

Переортогонализация поэтому не делает соотношение между b'_2 , $A c_1$ и c_1 сколько-нибудь менее удовлетворительным, но она дает такой вектор b'_2 , который ортогонален к c_1 с рабочей точностью в том смысле, что угол между ними отличается от $\pi/2$ на величину порядка 10^{-5} .

Это иллюстрируется в табл. 3. Нормированный вектор вычисляется из b'_2 , так что

$$k_2 c_2 = b'_2. \quad (32.7)$$

Так как c_1 и c_2 теперь почти ортогональны, b_3 действительно с рабочей точностью равен нулю. Заметим, что вычисленные величины удовлетворяют соотношениям

$$k_2 c_2 = A c_1 - h_{11} c_1 + O(10^{-5}), \quad (32.8)$$

$$b_3 = O(10^{-5}) = A c_2 - h_{21} c_1 - h_{22} c_2 + O(10^{-5}), \quad (32.9)$$

так что

$$A [c_1 \parallel c_2] = [c_1 \parallel c_2] \begin{bmatrix} h_{11} & h_{12} \\ k_2 & h_{22} \end{bmatrix} + O(10^{-5}), \quad (32.10)$$

что является наилучшим результатом, который можно было бы ожидать. Он показывает, что вычисленная H сейчас почти точно подобна A . Заметим, что след сохранился с рабочей точностью, и можно проверить, что это же верно и для собственных значений, а также для собственных векторов, получающихся после преобразования.

Значение переортогонализации

33. Для того чтобы получить b'_2 , который действительно ортогонален c_1 , мы применяли процесс ортогонализации Шмидта к b_2 . Важно понять, что поправка к b_2 , полученная таким путем, не единственная, дающая удовлетворительные результаты. Все, что нам нужно, — это получить такую поправку, которая была бы порядка не больше 10^{-5} и делала бы b'_2 почти ортогональным c_1 . Например, вектор

$$b'_2 = 10^{-4}(0,43814, -0,55133) \quad (33.1)$$

дает такие же хорошие результаты, как и b'_2 , полученный при помощи переортогонализации. После нормировки он дает такой c_2 , что (32.8) выполнено. Эффект взаимного уничтожения дает некоторую степень свободы в выборе вектора b'_2 . В работе Уилкинсона (1963b, стр. 86—91) показано, что если происходит взаимное уничтожение, даже при использовании переортогонализации окончательные матрицы C и H очень сильно зависят от точности вычислений.

Простейшая процедура, вообще говоря, состоит в следующем. Вычисляем вектор b_{r+1} вида

$$b_{r+1} = Ac_r - h_{1r}c_1 - h_{2r}c_2 - \dots - h_{rr}c_r. \quad (33.2)$$

Затем вычисляем b'_{r+1} такой, что

$$b'_{r+1} = b_{r+1} - \varepsilon_{1r}c_1 - \varepsilon_{2r}c_2 - \dots - \varepsilon_{rr}c_r, \quad (33.3)$$

$$h_{r+1}c_{r+1} = b_{r+1}, \quad \|c_{r+1}\|_2 = 1, \quad (33.4)$$

где

$$\varepsilon_{ir} = c_i^T b_{r+1} / c_i^T c_i = c_i^T b_{r+1}. \quad (33.5)$$

Тогда будем иметь

$$A(c_1 \| c_2 \dots \| c_n) = (c_1 \| c_2 \dots \| c_n)H + O(2^{-t}), \quad (33.6)$$

если A — нормированная матрица и мы используем вычисления с t разрядами. Точное значение $O(2^{-t})$ обдуманно оставлено неопределенным. Вообще говоря, это будет матрица, норма которой ограничена величиной $f(n)2^{-t}$, где $f(n)$ зависит от типа используемой арифметики. Множитель $f(n)$, естественно, должен быть отнесен к эффекту накопления ошибок округления; для вычислений с плавающей запятой и накоплением скалярных произведений грубая оценка показывает, что $f(n) < Kn^2$, но этот результат, вероятно, можно улучшить.

Необходимость переортогонализации возникает из-за взаимного уничтожения, а не из-за накопления ошибок округления. Для того чтобы подчеркнуть это, мы использовали матрицу второго порядка. Если на всех этапах не происходит взаимного уничтожения, то переортогонализацию можно опустить, но, к сожалению, такие случаи редки. Включение переортогонализации делает метод Арнольди значительно менее экономичным, чем метод Хаусхолдера; в нем $2n^3$ умножений без переортогонализации и $3n^3$ с переортогонализацией. (В этой связи смотрите сравнение процесса ортогонализации Шмидта и процесса приведения к треугольному виду Хаусхолдера в гл. 4, § 55). Без переортогонализации метод Арнольди может давать произвольно плохие результаты даже для хорошо обусловленных матриц. Однако если переортогонализация включена, то метод Арнольди сравним по точности с методом Хаусхолдера; *эта точность*

никоим образом не уменьшается, если случайно мы выберем b_i так, что он имеет малые или даже полностью не имеет компонент по некоторым собственным векторам.

Снова имеем ситуацию, когда при точных вычислениях, если b_r равен нулю, имеется полная свобода в выборе «всех знаков» c_r при условии лишь, что он ортогонален к предыдущим c_i . При вычислениях же с ограниченной точностью, когда k цифр исчезают в результате взаимного уничтожения, можем выбирать $t - k$ знаков произвольно. Заметим, что в методе Арнольди мы не можем иметь срыва второго типа, описанного в § 27, так как $x_i^T c_i = c_i^T c_i = 1$.

34. В связи с переортогонализацией весьма естественно возникают два вопроса:

(i) Почему в методе Хессенберга нет ничего соответствующего этой переортогонализации?

(ii) Что занимает место переортогонализации в приведениях Гивенса и Хаусхолдера, которые, как мы знаем, математически эквивалентны приведению Арнольди?

Необходимо, чтобы читателю все стало ясно в этих вопросах, и мы детально обсудим их. Обсудим сначала первый вопрос. Заметим, что если мы определим c_{r+1} уравнениями

$$h_{r+1}c_{r+1} = Ac_r - h_{1r}c_1 - \dots - h_{rr}c_r, \quad (34.1)$$

то, предполагая только, что $c_{n+1} = 0$, получим

$$AC = CH, \quad (34.2)$$

как бы ни были выбраны h_{ir} . Такой выбор h_{ir} , при котором c_{r+1} ортогонален к линейно независимым векторам x_i , важен только потому, что он обеспечивает линейную независимость c_i и равенство нулю c_{n+1} . Если x_i суть векторы e_i , точная ортогональность обеспечивается если взять соответствующие компоненты c_i точно равными нулю. Поэтому мы достигаем точной ортогональности, почти не заботясь о ней. При большинстве выборов x_i точная ортогональность не может быть достигнута таким простым способом, и обычно требует значительных вычислений. Однако если X — любая неособенная верхняя треугольная матрица, векторы c_i в точности те же самые, что и в случае $X = I$. Действительно, в этом случае, для того чтобы вектор c_{r+1} был ортогонален $x_1, \dots, x_r$, необходимо, чтобы его первые r компонент были точно равны нулю. При вычислении b_{r+1} из соотношений

$$b_{r+1} = Ac_r - h_{1r}c_1 - h_{2r}c_2 - \dots - h_{rr}c_r \quad (34.3)$$

мы поэтому никак не учитываем действительного значения компонент x_i . Сначала вычитаем c_1 , умноженный на h_{1r} , из Ac_r так, чтобы пропала первая компонента. Тогда мы имеем ортогональность к x_1 . Затем вычитаем c_2 , умноженный на h_{2r} так, чтобы пропала вторая компонента. При этом первая компонента остается равной нулю. Продолжая таким образом, мы получаем b_{r+1} , первые r компонент которого равны нулю и который поэтому ортогонален $x_1, \dots, x_r$.

Второй вопрос несколько более тонкий. Векторы c_i , полученные по методу Арнольди, являются аналогами столбцов матрицы, полученной перемножением преобразующих ортогональных матриц в методах Гивенса или Хаусхолдера. Мы видели в главе 3, что сравнительно простые приемы

обеспечивают ортогональность этих множителей с рабочей точностью. В методе Арнольди матрица преобразования не факторизована, и определяем ее прямо, столбец за столбцом. Так случилось, что численный процесс, используемый для получения ортогональности в этом случае значительно менее эффективный.

Метод Ланцоша

35. Еще один выбор системы x_i имеет важное практическое значение. В этом случае x_i определяются одновременно с b_i и c_i , и они получаются по A^T таким же образом, как c_i получаются по A . Сначала обсудим точный математический процесс. В связи с тем, что c_i и x_i играют столь тесно связанные роли, нам удобно использовать обозначения, которые отразили бы это.

Следуя Ланцошу (1950), будем обозначать векторы, полученные по A^T , через b_i^* и c_i^* . (Мы хотим подчеркнуть, что звездочка не обозначает комплексного сопряжения здесь и далее в этом параграфе). Соответствующие уравнения будут

$$k_{r+1}c_{r+1} = b_{r+1} = Ac_r - \sum_{i=1}^r h_{ir}c_i, \quad (35.4)$$

$$k_{r+1}^*c_{r+1}^* = b_{r+1}^* = A^Tc_r^* - \sum_{i=1}^r h_{ir}^*c_i^*, \quad (35.5)$$

где h_{ir} и h_{ir}^* выбраны так, что b_{r+1} ортогонален $c_1^*, \dots, c_r^*$, а b_{r+1}^* ортогонален $c_1, \dots, c_r$ соответственно.

К обоим последовательностям можно приложить общий анализ, и следовательно, имеем

$$\left. \begin{aligned} AC &= CH, & (C^*)^T C &= L, \\ A^T C^* &= C^* H^*, & C^T C^* &= L^*, \end{aligned} \right\} \quad (35.6)$$

где H и H^* — в верхней форме Хессенберга, а L и L^* — нижние треугольные. Из второго и четвертого уравнений (35.6) находим

$$L = (L^*)^T, \quad (35.7)$$

и, следовательно, обе эти матрицы диагональные и могут быть обозначены через D . Тогда первое и третье уравнения дают

$$H = C^{-1}AC = D^{-1}(H^*)^T D. \quad (35.8)$$

Слева в (35.5) стоит верхняя матрица Хессенберга, а справа — нижняя. Следовательно, обе они трехдиагональные, и

$$h_{ir}^* = h_{ir} = 0 \quad (i = 1, \dots, r-2), \quad (35.9)$$

откуда ясно, что если Ac_r ортогонален к c_{r-1}^* и c_r^* , то он автоматически ортогонален ко всем предыдущим c_i^* и аналогично для $A^Tc_r^*$. Далее из (35.5) имеем

$$h_{rr} = h_{rr}^*, \quad h_{r, r+1}k_{r+1} = h_{r, r+1}^*k_{r+1}^*, \quad (35.10)$$

так что если множители k_{r+1} и k_{r+1}^* выбраны равными единице, $h_{r, r+1} = h_{r, r+1}^*$ и H и H^* совпадают. Так как H и H^* — трехдиагональные,

ные, проще использовать обозначения

$$\gamma_{r+1}c_{r+1} = b_{r+1} = Ac_r - \alpha_r c_r - \beta_r c_{r-1}, \quad (35.8)$$

$$\gamma_{r+1}^* c_{r+1}^* = b_{r+1}^* = A^T c_r^* - \alpha_r c_r^* - \beta_r^* c_{r-1}^*, \quad (35.9)$$

где из (35.7)

$$\gamma_{r+1}\beta_{r+1} = \gamma_{r+1}^*\beta_{r+1}^*. \quad (35.10)$$

Удобно дать точные выражения для α_r , β_r , β_r^* . Имеем

$$\alpha_r = (c_r^*)^T A c_r / (c_r^*)^T c_r = c_r^T A^T c_r^* / c_r^T c_r^*, \quad (35.11)$$

$$\begin{aligned} \beta_r &= (c_{r-1}^*)^T A c_r / (c_{r-1}^*)^T c_{r-1} = c_r^T A^T c_{r-1}^* / (c_{r-1}^*)^T c_{r-1} = \\ &= c_r^T (\gamma_r^* c_r^* + \alpha_{r-1} c_{r-1}^* + \beta_{r-1}^* c_{r-2}^*) / (c_{r-1}^*)^T c_{r-1} = \gamma_r^* c_r^T c_r^* / (c_{r-1}^*)^T c_{r-1} \end{aligned} \quad (35.12)$$

из ортогональности и (35.9). Аналогично

$$\beta_r^* = \gamma_r (c_r^*)^T c_r / (c_{r-1})^T c_{r-1}^*. \quad (35.13)$$

Срыв процесса

36. Прежде чем рассматривать практический процесс, мы должны рассмотреть случай, когда либо b_{r+1} , либо b_{r+1}^* (или оба) равны нулю. Эти события независимы. Если b_{r+1} равен нулю, берем в качестве c_{r+1} произвольный вектор, ортогональный к c_1^* , ..., c_r^* , и (35.8) выполняется, если положить $\gamma_{r+1} = 0$. Аналогично, если b_{r+1}^* равен нулю, берем в качестве c_{r+1}^* произвольный вектор, ортогональный к c_1 , ..., c_r , и $\gamma_{r+1}^* = 0$. Заметим из (35.10), что если b_{r+1} равен нулю, а b_{r+1}^* нет, то β_{r+1}^* равен нулю, а если b_{r+1}^* равен нулю, а b_{r+1} нет, то β_{r+1} равен нулю.

Использование множителей γ_{r+1} , γ_{r+1}^* делает очевидным то обстоятельство, что обращение в нуль b_{r+1} или b_{r+1}^* не должно рассматриваться как «срыв» процесса. Оно просто дает нам свободу выбора. Если положим множители равными единице, то обращение в нуль b_{r+1} или b_{r+1}^* больше кажется исключительным, чем является им.

С другой стороны, если $c_r^T c_r^*$ равно нулю на каком-либо этапе, процесс полностью срывается, и мы должны начать его снова с другим вектором c_1 или c_1^* (или обоими) в надежде, что больше этого не случится. К сожалению, срыв по-видимому не дает каких-либо указаний для нового выбора c_1 или c_1^* .

Предположим, что c_1 и c_1^* таковы, что $c_i^T c_i^* \neq 0$ при всех соответствующих значениях i . Покажем, что если в этом случае c_1 высоты p по отношению к A , то первый b_i , который обратится в нуль, есть b_{p+1} .

Действительно, комбинируя уравнения (35.8) для $r = 1, \dots, s$, мы можем написать

$$b_{s+1} = (A^s + q_{s-1}^{(s)} A^{s-1} + q_{s-2}^{(s)} A^{s-2} + \dots + q_0^{(s)} I) c_1 = f_s(A) c_1, \quad (36.1)$$

где $f_s(A)$ — нормализованный полином (гл. 1, § 18) степени s от A . Следовательно, b_{s+1} не может обратиться в нуль, если $s < p$. Так как c_1 высоты p , $A^i c_1$ может быть записан в виде линейной комбинации $A^i c_1$ ($i = 1, 2, \dots, p-1$), так что

$$b_{p+1} = f_p(A) c_1 = g_{p-1}(A) c_1, \quad (36.2)$$

где $g_{p-1}(A)$ — полином степени $p-1$ от A . Из уравнений (36.1) для $s = 1, \dots, p-1$ мы видим, что $g_{p-1}(A) c_1$ может быть выражен в виде

линейной комбинации $b_p, \dots, b_1$ и, следовательно, $c_p, \dots, c_1$. Можно написать

$$b_{p+1} = \sum_{i=1}^p t_i c_i, \quad (36.3)$$

и так как b_{p+1} ортогонален к $c_p^*, \dots, c_1^*$, имеем

$$0 = t_i (c_i^*)^T c_i, \quad \text{т. е.} \quad t_i = 0 \quad (i = 1, \dots, p), \quad (36.4)$$

и доказательство завершено.

Мы получили, что если A неполная, то мы должны иметь преждевременное обращение в нуль b_i и соответственно b_i^* .

Численный пример

37. Мы хотим подчеркнуть, что обращение в нуль $c_i^T c_i^*$ не связано с какими-либо плохими свойствами матрицы A . Это может случиться, даже если проблема собственных значений очень хорошо обусловлена. Мы вынуждены рассматривать это как специфическую слабость самого метода Ланцоша. В табл. 4 дается пример такого срыва. Так как использование нормирующих множителей вводит ошибки округления, мы брали все γ_i и γ_i^* равными единице, так что $c_i = b_i$ и $c_i^* = b_i^*$ на всех стадиях. Величина $c_2^T c_2^*$ в точности равна нулю. Очевидно, что это будет верно, какие бы множители ни использовались при предположении, что не делается ошибок округления. Для матриц, не имеющих малых целых элементов, если взять начальные векторы со случайными компонентами, точное обращение в нуль $c_i^T c_i^*$ абсолютно невероятно.

Т а б л и ц а 4

$$A = \begin{bmatrix} 5 & 1 & -1 \\ -5 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad c_1 = \begin{bmatrix} 0,6 \\ -1,4 \\ 0,3 \end{bmatrix}, \quad c_1^* = \begin{bmatrix} 0,6 \\ 0,3 \\ -0,1 \end{bmatrix},$$

$$c_2 = \frac{1}{3} \begin{bmatrix} 1,5 \\ -2,5 \\ 1,5 \end{bmatrix}, \quad c_2^* = \frac{1}{3} \begin{bmatrix} 1,8 \\ 0,6 \\ -0,8 \end{bmatrix}, \quad c_2^T c_2^* = 0.$$

Практический процесс Ланцоша

38. Если процесс Ланцоша проводится на практике, причем используются уравнения (35.8)–(35.13), то строгая биортогональность обычно быстро утрачивается. Если при образовании векторов c_i или c_i^* происходит значительное взаимное уничтожение, то немедленно происходит катастрофическое ухудшение ортогональности. Для того чтобы избежать этого, существенно, чтобы переортогонализация выполнялась на каждой стадии. Удобная практическая процедура заключается в следующем.

На каждой стадии мы сначала вычисляем b_{r+1} и b_{r+1}^* вида

$$\left. \begin{aligned} b_{r+1} &= A c_r - \alpha_r c_r - \beta_r c_{r-1}, & \alpha_r &= (c_r^*)^T A c_r / (c_r^*)^T c_r, \\ & & \beta_r &= (c_{r-1}^*)^T A c_r / (c_{r-1}^*)^T c_{r-1}, \\ b_{r+1}^* &= A^T c_r^* - \alpha_r^* c_r^* - \beta_r^* c_{r-1}^*, & \alpha_r^* &= c_r^T A^T c_r^* / (c_r^*)^T c_r, \\ & & \beta_r^* &= (c_{r-1}^*)^T A^T c_r^* / (c_{r-1}^*)^T c_{r-1}. \end{aligned} \right\} \quad (38.1)$$

Полезно определять два значения α_r независимо для проверки вычислений.

Затем определяются c_{r+1} и c_{r+1}^* из соотношений

$$\left. \begin{aligned} \gamma_{r+1} c_{r+1} &= \bar{b}_{r+1} = b_{r+1} - \varepsilon_{r1} c_1 - \varepsilon_{r2} c_2 - \dots - \varepsilon_{r1} c_r, \\ \gamma_{r+1} c_{r+1}^* &= \bar{b}_{r+1}^* = b_{r+1}^* - \varepsilon_{r1}^* c_1^* - \varepsilon_{r2}^* c_2^* - \dots - \varepsilon_{rr}^* c_r^*, \end{aligned} \right\} \quad (38.2)$$

где

$$\varepsilon_{ri} = (c_i)^T b_{r+1} / (c_i)^T c_i, \quad \varepsilon_{ri}^* = c_i^T b_{r+1}^* / c_i^T c_i^*. \quad (38.3)$$

Другими словами, b_{r+1} переортогонализуется по отношению к $c_1^*, \dots, c_r^*$, а b_{r+1}^* — по отношению к $c_1, \dots, c_r$. Естественно, что это осуществляется последовательно, причем каждая поправка вычитается сразу после того, как она вычислена. Нельзя категорически утверждать, что переортогонализация не дает векторов, которые были бы получены при помощи более точных вычислений, и что она не имеет вредных эффектов.

Заметим, что одинаково важно переортогонализировать b_{r+1} как по отношению к c_{r-1}^*, c_r^* , так и по отношению к $c_1^*, \dots, c_{r-2}^*$. Это происходит потому, что необходимость переортогонализации возникает из взаимного уничтожения, а не из-за накопления ошибок округления. Множители γ_{r+1} и γ_{r+1}^* могут быть выбраны степенями двух так, чтобы максимальные элементы $|c_{r+1}|$ и $|c_{r+1}^*|$ заключались между $1/2$ и 1 , так как это не вносит ошибок округления. При использовании вычислений с плавающей запятой эти нормирующие множители могут не иметь значения, но они становятся важными при вычислении собственных векторов полученной трехдиагональной матрицы (см. гл. 5, § 67).

Мы знаем, что теоретически процесс срывается, если на какой-либо стадии $c_i^T c_i^* = 0$. Поэтому мы можем ожидать, что вычислительный процесс будет неустойчивым, если $c_i^T c_i^*$ мало. (Мы здесь предполагаем, что c_i и c_i^* нормированы.) Если все $c_i^T c_i^*$ порядка единицы, то можно показать, что для нормированной A (35.5) дает

$$AC = CT + O(2^{-i}), \quad (38.4)$$

где T это трехдиагональная матрица с элементами

$$t_{ii} = \alpha_i, \quad t_{i, i+1} = \beta_{i+1}, \quad t_{i+1, i} = \gamma_{i+1}. \quad (38.5)$$

Можно заметить, что мы могли бы заменить T на $(T + F)$, где F — матрица, образованная ε_{ij} , и где $(T + F)$ в верхней форме Хессенберга с малыми элементами вне трехдиагональной секции. Но нет никакого смысла делать это, так как вклад F уже учтен членом $O(2^{-i})$. Точная оценка для $O(2^{-i})$ зависит от типа использованной арифметики.

Однако если

$$c_i^T c_i^* \approx 2^{-p} \quad (p — \text{положительное целое число}) \quad (38.6)$$

при некотором i , то, вообще говоря,

$$AC = CT + O(2^{2p-i}), \quad (38.7)$$

так что обычно мы теряем в точности на $2p$ знаков больше, чем в случае, когда A преобразуется с использованием устойчивой процедуры.

Численный пример.

39. Природа такой неустойчивости хорошо иллюстрируется примером в табл. 5, в котором для простоты мы положили c_1 и c_1^* почти ортогональными. (Мы можем избежать срыва на первом шаге на практике, положив $c_1 = c_1^*$). Вычисления проведены с высокой точностью, так что все данные

величины точны во всех знаках. Мы выбрали γ_i и γ_i^* равными единице для того, чтобы проиллюстрировать наше утверждение. При этом β_i и β_i^* равны.

Заметим, что если мы обозначим $c_1^T c_1^*$ через ε , то α_1 и α_2 будут порядка ε^{-1} и почти равны и противоположны по знаку, а β_2 — порядка ε^{-2} . При помощи довольно скучного анализа можно показать, что, вообще говоря, если $c_r^T c_r = \varepsilon$, то α_r и α_{r+1} порядка $\|A\|/\varepsilon$, β_{r+1} порядка $\|A\|^2/\varepsilon^2$, а β_{r+2} порядка $\|A\|^2\varepsilon$. (Последний результат не виден из примера в табл. 5, так как для этого нужны значения ε , значительно меньшие 2^{-9} .) Рост, вызванный малостью $c_r^T c_r$, затем прекращается, если только некоторые из последующих $c_i^T c_i^*$ не малы, но это будет, конечно, независимым событием.

Непосредственно можно видеть, почему значение величины $c_r^T c_r^*$ порядка 2^{-p} вызывает эффект, указанный в (38.7). Предположим, например, что мы имеем дело только с матрицей второго порядка, и для α_1 , α_2 и β_2 получены правильно округленные точные значения. Вычисленная матрица может быть тогда представлена в виде

$$\begin{bmatrix} \alpha_1(1 + \varepsilon_1) & \beta_2(1 + \varepsilon_2) \\ 1 & \alpha_2(1 + \varepsilon_3) \end{bmatrix}, \quad |\varepsilon_i| \leq 2^{-t}, \quad (39.1)$$

где α_i и β_i — значения, соответствующие точным вычислениям. Характеристический полином матрицы из (39.1) равен

$$\lambda^2 - \lambda[\alpha_1(1 + \varepsilon_1) + \alpha_2(1 + \varepsilon_3)] + \alpha_1\alpha_2(1 + \varepsilon_1)(1 + \varepsilon_3) - \beta_2(1 + \varepsilon_2). \quad (39.2)$$

Следовательно, ошибка в постоянном члене равна

$$\alpha_1\alpha_2(\varepsilon_1 + \varepsilon_3 + \varepsilon_1\varepsilon_3) - \beta_2\varepsilon_2, \quad (39.3)$$

и мы видели, что α_1 и α_2 порядка 2^p , а β_2 порядка 2^{2p} . Нет никаких причин для того, чтобы члены в (39.3) взаимно уничтожились, и, вообще говоря, этого не происходит. Этот пример показал, что трудности возникают не из-за накопления ошибок предыдущих шагов.

Если бы мы использовали обычные нормирующие множители в примере в табл. 5, то вместо матрицы T мы бы получили матрицу $\tilde{T}$, однако все наши выводы не зависят от этого. «Убыток» все равно пропорционален 2^{2p} , а не 2^p .

Таблица 5

$$\begin{array}{ccc} A & c_1 & c_1^* \\ \begin{bmatrix} 4 & 1 & 3 & 2 \\ 2 & 1 & 2 & 5 \\ 1 & 3 & 3 & 4 \\ 4 & 1 & 2 & 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 2^{-10} \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 2^{-10} \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad c_1^T c_1^* = 2^{-9}$$

$$T = \begin{bmatrix} 1026,5005 & -1037297,7 & -1,1047112 & 1,0358765 \\ 1 & -1010,3982 & -8,7263879 & 1,6241315 \\ & 1 & 1 & 1 \end{bmatrix}$$

$$\tilde{T} = \begin{bmatrix} 1026,5005 & -1037,2977 & -1,1047112 & 1,0358765 \\ 1000 & -1010,3982 & -8,7263879 & 1,6241315 \\ & 1 & 1 & 1 \end{bmatrix}$$

Процесс Ланцоша проводился для той же матрицы с использованием начальных векторов

$$c_1^T = (1, 2^{-p}, 0, 0), \quad (c_1^*)^T = (2^{-p}, 1, 0, 0)$$

с различными значениями p , причем использовались вычисления с двойной точностью (60 двоичных знаков). С возрастанием p собственные значения вычисленных T все дальше отдалялись от собственных значений A , и при $p = 30$ они не имели ни одной общей значащей цифры, что подтверждает наш анализ. Порядки величин $\alpha_1, \alpha_2, \beta_2, \beta_3$ изменялись так, как описано ранее в этом параграфе.

Общие замечания к несимметричному процессу Ланцоша

40. Хотя процесс Ланцоша дает нам трехдиагональную матрицу, т. е. очень сильное сокращение исходных данных, это может потребовать высокой платы в смысле численной устойчивости. Из-за этой потенциальной неустойчивости процесс Ланцоша был запрограммирован с использованием арифметики с двойной точностью и с указанием допустимых значений для $|c_i^T c_i^*|$ при нормированных c_i и c_i^* . Если модуль этого скалярного произведения оказывался меньше $2^{-\frac{1}{2}t}$, процесс вновь начинался с другими начальными векторами. Хотя такое повторение процесса случалось редко, умеренно малые значения $c_i^T c_i^*$ довольно обычны; это показывает, что точность, достижимая при вычислениях с одинарной точностью, часто бывает недостаточной.

Полное число умножений с включением переортогонализации приблизительно равно $4n^3$, и так как всюду проводится работа с двойной точностью, общий объем вычислений весьма значителен. Заметим, однако, что исходная матрица дается с обычной точностью, и следовательно, при образовании Ac_i и $A^T c_i^*$ (что всегда включает $2n^3$ умножений) нам нужно только перемножать элементы A , заданные с обычной точностью, на элементы c_i и c_i^* , заданные с двойной точностью. Мы покажем в § 49, что существует другой процесс, который в смысле устойчивости является полным эквивалентом процесса Ланцоша, и требует значительно меньшего объема работы. Ввиду этого мы не будем больше обсуждать несимметричный процесс Ланцоша.

Симметричный процесс Ланцоша

41. Если A — вещественная и симметричная, и мы берем $c_1 = c_1^*$, то наши две последовательности векторов совпадают. Если, кроме того, на каждой стадии мы выбираем γ_{r+1} так, что $\|c_{r+1}\|_2^2 = 1$, то видно, что симметричный процесс Ланцоша и симметричный процесс Арнольди совпадают. В симметричном процессе Ланцоша переортогонализация так же важна, как и в несимметричном, но требует вдвое меньше вычислений. Однако, как уже отмечалось в связи с общим случаем метода Арнольди, мы не можем иметь $c_i^T c_i^* = 0$, так как $c_i^T c_i^* = \|c_i\|_2^2 = 1$ из нормировки.

Если переортогонализация включена, то численная устойчивость симметричного метода Ланцоша сравнима с численной устойчивостью симметричного приведения Гивенса и Хаусхолдера к трехдиагональному виду, но, конечно, приведение Хаусхолдера значительно более экономно. Никакой потери точности не будет, если начальный вектор имеет малые

или не имеет совсем составляющих по некоторым собственным векторам, но переортогонализация будет особенно важна в этих случаях. Если мы возьмем $c_1 = e_1$, то с точностью до знаков процесс Ланцоша дает ту же самую трехдиагональную матрицу, что и методы Гивенса и Хаусхолдера, при точных вычислениях. Трудно придумать какую-либо причину, по которой мы использовали бы метод Ланцоша в предпочтение методу Хаусхолдера.

Мы предполагали, что нормирующие множители γ_r выбраны так, что на каждой стадии $\|c_r\|_2^2 = 1$, так как это делает связь с другими методами особенно тесной. Однако этот выбор не служит никакой другой полезной цели и вызывает необходимость использования квадратных корней. На практике проще выбирать γ_r степенями двух, так чтобы наибольший элемент $|c_r|$ лежал между $1/2$ и 1 . При точных вычислениях все β_r неотрицательны, так как из (35.12) имеем

$$\beta_r = \gamma_r c_r^T c_r / c_{r-1}^T c_{r-1} = \gamma_r \|c_r\|^2 / \|c_{r-1}\|^2. \quad (41.1)$$

Однако β_r вычисляются не таким способом, и возможно, что вычисленные β_r будут малыми отрицательными величинами. В этом случае T не будет квазисимметричной (гл. 5, § 66). При вычислении собственных значений T такие β_r должны быть заменены нулями.

Приведение матриц Хессенберга к более компактному виду

42. Итак, существует несколько очень устойчивых методов приведения матриц общего вида к форме Хессенберга. Остаток этой главы посвящен дальнейшему приведению матриц Хессенберга к некоторым более компактным формам. Обсудим две такие формы — трехдиагональную форму и форму Фробениуса. В связи с несимметричным процессом Ланцоша уже рассматривался вопрос о приведении к трехдиагональной матрице и отмечалось, что это возможно только с риском численной неустойчивости. Это же будет верно, вообще говоря, для всех методов, которые обсуждаются в оставшейся части главы.

Приведение нижней матрицы Хессенберга к трехдиагональному виду

43. Рассмотрим сначала приложение метода § 11 к матрице A в нижней форме Хессенберга, опуская на время вопрос перестановок. Следовательно, нам нужно найти единичную нижнюю треугольную матрицу N , первый столбец которой равен e_1 , такую, что

$$AN = NH, \quad (43.1)$$

где H — в верхней форме Хессенберга. Покажем, что в этом случае H должна быть трехдиагональной. Нужно только показать, что $h_{ir} = 0$ ($r > i + 1$), так как мы знаем, что H — верхняя матрица Хессенберга.

Будем последовательно приравнивать элементы r -го столбца в обеих частях уравнения (43.1). Так как A — нижняя матрица Хессенберга, а N — нижняя треугольная, (i, r) -элемент AN равен нулю, если $i <$

$< r - 1$. Таким образом, мы последовательно имеем:

$$\left. \begin{aligned} 0 &= h_{1r}, \\ 0 &= h_{2r}, \\ 0 &= n_{32}h_{2r} + h_{3r} = h_{3r}, \\ 0 &= n_{42}h_{2r} + n_{43}h_{3r} + h_{4r} = h_{4r}, \\ &\dots\dots\dots \\ 0 &= n_{r-2,2}h_{2r} + n_{r-2,3}h_{3r} + \dots + h_{r-2,r} = h_{r-2,r}, \end{aligned} \right\} \quad (43.2)$$

и результат доказан.

Для доказательства несущественно, что первый столбец N равен e_1 . Если мы возьмем в качестве первого столбца N произвольный вектор с первым элементом, равным единице, то мы просто введем в каждое из уравнений (43.2), кроме первого, член $n_{i1}h_{1r}$. Матрицы N и H определятся целиком, и H все равно будет трехдиагонального вида.

Для отличных от нуля элементов r -го столбца H имеем:

$$\left. \begin{aligned} a_{r-1,r} &= h_{r-1,r}, \\ a_{rr} + a_{r,r+1}n_{r+1,r} &= n_{r,r-1}h_{r-1,r} + h_{rr}, \\ a_{r+1,r} + a_{r+1,r+1}n_{r+1,r} + a_{r+1,r+2}n_{r+2,r} &= \\ &= n_{r+1,r-1}h_{r-1,r} + n_{r+1,r}h_{rr} + h_{r+1,r}, \end{aligned} \right\} \quad (43.3)$$

а для отличных от нуля элементов $(r+1)$ -го столбца N —

$$\begin{aligned} a_{ir} + a_{i,r+1}n_{r+1,r} + a_{i,r+2}n_{r+2,r} + \dots + a_{i,i+1}n_{i+1,r} = \\ = n_{i,r-1}h_{r-1,r} + n_{ir}h_{rr} + n_{i,r+1}h_{r+1,r} \quad (i = r+2, \dots, n). \end{aligned} \quad (43.4)$$

Если какой-либо индекс больше n , соответствующий член опускается. Скалярные произведения могут накапливаться, как в § 11, но соотношения упрощаются в силу специальных форм A и H . Весь процесс требует приблизительно $\frac{1}{6} n^3$ умножений.

Использование перестановок

44. Мы описали процесс, аналогичный процессу § 11, а не § 8, так как первый более экономичен при рассмотрении ошибок округления. При рассмотрении перестановок несколько более удобно вернуться к § 8. Если A_0 задана как нижняя форма Хессенберга, то при предположении, что перестановки не использовались, A_{r-1} имеет, например, для случая $n = 6$, $r = 3$ вид

$$\begin{bmatrix} \times & \times & & & & \\ \times & \times & \times & & & \\ 0 & \times & \times & \times & & \\ 0 & 0 & \times & \times & \times & \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \end{bmatrix}. \quad (44.1)$$

Основной r -й шаг заключается в следующем.

Для всех значений i от $r+2$ до n выполняем операции (i) и (ii):

(i) Вычисляем $n_{i,r+1} = a_{ir}/a_{r+1,r}$.

(ii) Вычитаем $n_{i,r+1} \times$ (строка $(r+1)$) из строки i . (Заметим, что строка $(r+1)$ имеет только два отличных от нуля элемента, кроме $a_{r+1,r}$.)

(iii) Для всех значений i от $r + 2$ до n прибавляем $n_{i, r+1} \times$ (столбец i) к столбцу $r + 1$.

Очевидно, что r -й шаг не портит ни один из нулей первоначальной A_0 или введенных на предыдущих шагах.

Однако если $a_{r+1, r} = 0$ и $a_{i, r} \neq 0$ при некотором i , то процесс срывается на r -м шаге. Вообще говоря, мы можем избежать срыва такого рода при помощи перестановок, но если мы переставим строки $(r + 1)$ и $(r + 1)'$ и столбцы $(r + 1)$ и $(r + 1)'$ обычным образом, мы нарушим систему нулей над диагональю. *Перестановки несовместимы с сохранением исходной нижней формы Хессенберга.*

Единственная свобода выбора состоит в том, что мы можем выполнить подобное преобразование с произвольной матрицей типа N_1 перед тем, как начнем приведение к трехдиагональному виду. Если произойдет срыв, мы можем вернуться к началу процесса и выполнить такое преобразование в надежде, что срыва больше не будет. В § 51 мы покажем, как слегка улучшить эту весьма случайную процедуру.

Так как перестановки не разрешаются, то при получении трехдиагональной матрицы существует очевидная опасность численной неустойчивости. Если на какой-либо стадии $|a_{r+1, r}|$ значительно меньше некоторых $|a_{i, r}|$ ($i = r + 2, \dots, n$), то соответствующий $|n_{i, r+1}|$ будет значительно больше единицы.

Легко видеть, что использование плоских вращений или матриц отражения также разрушает нижнюю форму Хессенберга, и поэтому не существует простого аналога, использующего ортогональные матрицы. Вообще говоря, мы нашли, что если «стабилизация» не разрешается в методах, использующих элементарные матрицы, то не существует соответствующих методов, использующих ортогональные матрицы.

Влияние малого ведущего элемента

45. Рассмотрим более детально эффект появления малого ведущего элемента $a_{r+1, r}^{(r-1)}$ в начале r -го шага. Для удобства обозначим этот элемент ε , а соответствующие элементы столбцов r , $r + 1$ и $r + 2$ через x_i , y_i , z_i . При $n = 7$, $r = 3$ матрица имеет вид

$$\begin{bmatrix} \times & \times & & & & & \\ \times & \times & \times & & & & \\ 0 & \times & \times & \times & & & \\ 0 & 0 & \varepsilon & y_4 & z_4 & & \\ 0 & 0 & x_5 & y_5 & z_5 & \times & \\ 0 & 0 & x_6 & y_6 & z_6 & \times & \times \\ 0 & 0 & x_7 & y_7 & z_7 & \times & \times \end{bmatrix}. \quad (45.1)$$

При r -м преобразовании меняются только выделенные столбцы. Для асимптотического поведения модифицированных элементов имеем

$$\left. \begin{aligned} a_{r+1, r+1}^{(r)} &= x_{r+2} z_{r+1} / \varepsilon + O(1), \\ a_{i, r+1}^{(r)} &= -x_i x_{r+2} z_{r+1} / \varepsilon^2 + O\left(\frac{1}{\varepsilon}\right) & (i = r + 2, \dots, n), \\ a_{i, r+2}^{(r)} &= -x_i z_{r+1} / \varepsilon + O(1) & (i = r + 2, \dots, n). \end{aligned} \right\} \quad (45.2)$$

Заметим, что элементы новых столбцов $(r+1)$ и $(r+2)$ находятся почти в прежних отношениях. Вообще множители на следующем шаге равны

$$n_{i, r+2} = a_{i, r+1}^{(r)} / a_{r+2, r+1}^{(r)} = x_i / x_{r+2} + O(\epsilon) \quad (45.3)$$

и, следовательно, будут порядка единицы. При умножении слева $A^{(r)}$ на N_{r+2} i -й элемент $(r+2)$ -го столбца будет

$$a_{i, r+2}^{(r)} - n_{i, r+2} a_{r+2, r+2}^{(r)} = -x_i z_{r+1} / \epsilon + O(1) - \\ - [x_i / x_{r+2} + O(\epsilon)] [-x_{r+2} z_{r+1} / \epsilon + O(1)] = O(1). \quad (45.4)$$

Этот простой анализ недооценивает степень взаимного уничтожения при выполнении преобразования подобия с матрицей N_{r+2} . Более детальный анализ показывает, что

$$a_{i, r+2}^{(r+1)} = O(\epsilon), \quad (45.5)$$

однако матрица N_{r+3} ни в коей мере не является исключительной (если не возникает независимой неустойчивости), и неустойчивость кончается. Ее главное влияние в окончательной трехдиагональной матрице сказывается поэтому в строках $(r+1)$, $(r+2)$, $(r+3)$, где элементы имеют следующие порядки величин:

$$\left. \begin{array}{lll} \text{строка } (r+1) & \epsilon & A/\epsilon \quad B \\ \text{строка } (r+2) & C/\epsilon^2 - A/\epsilon & D \\ \text{строка } (r+3) & E\epsilon & F \quad G \end{array} \right\}. \quad (45.6)$$

Здесь $A, B, \dots$ порядка $\|A\|$, и

$$A^2 \approx -BC. \quad (45.7)$$

Заметим, что расположение исключительных элементов в окончательной матрице является транспонированным по отношению к расположению в методе Ланцоша, если $c_{r+1}^T c_{r+1}^* = \epsilon$. В § 48 покажем, почему существует такая тесная связь.

Анализ ошибок

46. Очень детальный анализ ошибок такого приведения к трехдиагональному виду был проведен Уилкинсоном (1962a). Здесь ограничимся суммированием результатов. В таком анализе есть некоторая трудность. В § 14 мы видели, что даже при приведении матрицы общего вида к верхней форме Хессенберга, *при которой перестановки разрешены*, невозможно безусловно гарантировать устойчивость. То же самое происходит при приведении верхних матриц Хессенберга к трехдиагональному виду даже в наиболее благоприятных случаях, когда все множители порядка единицы. Однако не это будет предметом изучения сейчас. В основном займемся теми ошибками, которые возникают из-за появления малых ведущих элементов и связанного с этим появления больших множителей. В анализе Уилкинсона (1962a) мы поэтому пренебрегли всеми усиливающими эффектами, которые *могут быть* вызваны умножениями *нормальных* ошибок округления слева и справа на последовательность N_r и N_r^{-1} с элементами порядка единицы. При таких условиях результаты состоят в следующем.

Когда на нестабильной стадии возникают множители порядка 2^h , то эквивалентное возмущение некоторых элементов исходной матрицы A имеет порядок 2^{2h-t} , и, вообще говоря, это ведет к возмущению собствен-

ных значений, которое также порядка 2^{2k-t} . Если мы хотим получить результаты, сравнимые по точности с теми, которые получились бы, если бы в приведении не было неустойчивой стадии, нам нужно работать с $t + 2k$ значащими цифрами вместо t . Обратный анализ показывает, что такая точность необходима лишь для части вычислений, но на автоматических вычислительных машинах не так легко извлечь преимущества этого обстоятельства. *Анализ показал, что влияние ошибок, сделанных до неустойчивой стадии, не усиливается большими множителями.*

Общий порядок величин элементов окончательной трехдиагональной матрицы T в случае $n = 6$ с неустойчивостью на второй стадии таков:

$$|T| = \begin{bmatrix} 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ & 2^{-k} & 2^k & 1 & & \\ & & 2^{2k} & 2^k & 1 & \\ & & & 2^{-k} & 1 & 1 \\ & & & & 1 & 1 \end{bmatrix}. \quad (46.1)$$

Порядок величин элементов x и y , левого и правого собственных векторов T , соответствующих хорошо обусловленному собственному значению, равен

$$(1, 1, 2^k, 1, 1, 1), \quad (46.2)$$

$$(1, 1, 1, 2^k, 1, 1) \quad (46.3)$$

соответственно, где векторы нормированы так, что $y^T x$ порядка единицы. Соответствующее число обусловленности порядка 2^{2k} , так что имеет место значительное ухудшение. Однако здесь есть небольшой обман, так как матрица T плохо уравновешена. Она может быть преобразована при помощи диагонального преобразования подобия в матрицу $\tilde{T}$, элементы которой имеют следующие порядки величин:

$$|\tilde{T}| = \begin{bmatrix} 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ & 2^{-k} & 2^k & 2^k & & \\ & & 2^k & 2^k & 2^{-k} & \\ & & & 1 & 1 & 1 \\ & & & & 1 & 1 \end{bmatrix}, \quad (46.4)$$

и ошибок округления можно избежать, если в качестве элементов диагональной матрицы использовать степени двух. Соответствующие y и x снова такие, что $y^T x$ порядка единицы, и имеют следующие порядки величин:

$$(1, 1, 2^k, 2^k, 1, 1), \quad (46.5)$$

$$(1, 1, 1, 1, 1, 1); \quad (46.6)$$

число обусловленности теперь порядка 2^k . Тем не менее для вычисления собственных значений T или $\tilde{T}$ необходимо работать с на $2k$ большим числом значащих цифр, чем при работе с трехдиагональными матрицами, полученными без неустойчивой стадии.

Эти результаты показывают, что если мы готовы работать с двойной точностью, то можно придерживаться одной из следующих процедур.

(i) Мы можем согласиться допускать множители вплоть до порядка $2^{-t/2}$, и тогда, предполагая, что нет кратной неустойчивости (см. Уилкинсон (1962a)), получим ошибки, эквивалентные возмущению A , не больше

чем $2^{-2t} \times (2^{\frac{1}{2}t})^2$, т. е. 2^{-t} . Если множитель A превысит указанный допуск, то мы возвращаемся к матрице A , выполняем начальное преобразование подобия с некоторой матрицей N_1 , а затем повторяем приведение.

(ii) Если мы не хотим возвращаться к началу, то мы должны заменить все ведущие элементы, меньшие $2^{-2t/3}$, на $2^{-2t/3}$. При этом ошибки, вызванные этой заменой, будут сами порядка $2^{-2t/3}$, а ошибки при последующем приведении будут порядка $2^{-2t} (2^{2t/3})^2$, т. е. $2^{-2t/3}$. Поэтому это оптимальный выбор. Заметим, что если мы собираемся работать с тройной точностью, соответствующая граница будет 2^{-t} , и эквивалентное возмущение A будет по порядку не больше 2^{-t} .

Процесс Хессенберга в применении к нижней матрице Хессенберга

47. Мы установили в § 30, что если возьмем в § 26 $X = I$, то процесс Хессенберга совпадает, даже в смысле ошибок округления, с прямым приведением без перестановок. Так как это верно в общем случае, в частности это верно, если исходная матрица взята в нижней форме Хессенберга. Несмотря на это соотношение, полезно описать приведение снова в терминах процесса Хессенберга.

Перед началом r -го шага уже вычислены $c_1, c_2, \dots, c_r$, причем первые $(i - 1)$ элементов c_i равны нулю, а его i -й элемент равен единице. На r -м шаге вычисляем c_{r+1} из соотношения

$$k_{r+1}c_{r+1} = Ac_r - h_{1r}c_1 - h_{2r}c_2 - \dots - h_{rr}c_r, \quad (47.1)$$

где h_{ir} определены так, чтобы c_{r+1} был ортогонален к $e_1, \dots, e_r$. Если A — нижняя матрица Хессенберга, первые $(r - 2)$ элемента Ac_r равны нулю, и поэтому Ac_r уже ортогонален $e_1, \dots, e_{r-2}$; следовательно, $h_{ir} = 0$ ($i = 1, \dots, r - 2$). С точностью до множителя k_{r+1} , c_{r+1} определяется по Ac_r вычитанием c_r , умноженного на h_{rr} , и c_{r-1} , умноженного на $h_{r-1,r}$, так, чтобы $(r - 1)$ -й и r -й элементы стали равными нулю. Множитель k_{r+1} выбран так, что $(r + 1)$ -й элемент c_{r+1} равен единице. Очевидно, что матрица H трехдиагональная. Если вектор справа в (47.1) не равен нулю, но его $(r + 1)$ -й элемент равен нулю, то процесс срывается. Если весь вектор равен нулю, то в качестве c_{r+1} можно взять любой вектор, ортогональный к $e_1, \dots, e_r$, $(r + 1)$ -й элемент которого равен единице.

Если $c_1 = e_1$, векторы c_i совпадают со столбцами матрицы N в процессах §§ 43, 44. Выбор в качестве c_1 любого другого вектора, первая компонента которого равна единице, эквивалентен, в терминах § 44, включению начального преобразования $N_1 A N_1^{-1}$, где первый столбец N_1^{-1} равен c_1 .

Связь процесса Хессенберга и процесса Ланцоша

48. Сейчас покажем, что если A — нижняя матрица Хессенберга, то существует тесная связь процесса Хессенберга и несимметричного процесса Ланцоша.

Сначала докажем, что если в процессе Ланцоша мы возьмем $c_1^* = e_1$, то векторы c_i^* образуют столбцы верхней треугольной матрицы, незави-

симо от выбора c_1 . Действительно, предположим, что это верно вплоть до c_r^* . Имеем

$$\gamma_{r+1}^* c_{r+1}^* = A^T c_r^* - \alpha_r c_r^* - \beta_r^* c_{r-1}^*. \quad (48.1)$$

Но так как A^T — верхняя матрица Хессенберга, компоненты $A^T c_r^*$ в позициях $r+2, \dots, n$ равны нулю по нашей индуктивной гипотезе о c_r^* . Следовательно, компоненты c_{r+1}^* в этих позициях также равны нулю, независимо от возможных значений α_r и β_r^* . Это требуемый результат для c_{r+1}^* .

Мы можем использовать этот результат для того, чтобы показать, что векторы c_i являются столбцами нижней треугольной матрицы. Это следует из замечания в § 34 о том, что ортогонализация по отношению к столбцам любой верхней треугольной матрицы дает тот же результат, что и ортогонализация по отношению к столбцам I . Кроме того, там мы указали, что эта ортогонализация не использует элементы данной верхней треугольной матрицы.

Следовательно, векторы c_i и величины α_i и β_i в процессе Ланцоша можно получать без вычисления c_i^* . Так как они совпадают с величинами, которые получаются, если вместо $X = C^*$ взять $X = I$, мы установили совпадение векторов Ланцоша с векторами Хессенберга. Заметим, что в этом частном случае процесса Ланцоша не требуется переортогонализация, так как если X — верхняя треугольная, точная ортогональность достигается автоматически. Если, кроме того, $c_1 = e_1$, то и процесс Хессенберга, и процесс Ланцоша дают те же самые результаты, что и метод § 43.

Теперь видим, что совсем не удивительно, что возникновение малого ведущего элемента в процессе § 43 ведет к тому же поведению в трехдиагональной матрице, которое наблюдалось в § 39 для процесса Ланцоша в случае, когда $c_i^T c_i^*$ было мало. Это поведение является частным случаем общего явления, связанного с процессом Ланцоша.

Приведение матрицы общего вида к трехдиагональному виду

49. Можно скомбинировать методы §§ 12 и 43 так, чтобы получить приведение матрицы общего вида A к трехдиагональному виду. Сначала используем метод § 12 для приведения A к верхней форме Хессенберга H , используя перестановки. Он устойчив за исключением очень специальных случаев, и мы можем использовать арифметику с обычной точностью. Матрица H^T имеет нижнюю форму Хессенберга, и мы можем использовать метод § 43 (или эквивалентный ему метод § 47) для получения такой трехдиагональной матрицы T , что

$$H^T N = NT, \quad N^T H = T^T N^T. \quad (49.1)$$

Следовательно, трехдиагональная матрица T^T подобна A . Второе преобразование значительно менее устойчиво, и мы должны использовать арифметику с двойной точностью и учесть предосторожности, описанные в § 46.

Приведение к трехдиагональному виду требует всего $\frac{5}{6} n^3$ умножений

с обычной точностью и $\frac{1}{6} n^3$ умножений с двойной точностью. По очевидной

причине вторую часть процесса мы описали в терминах H^T , но так как $N^T = M$, единичной верхней треугольной матрице, первая строка которой

равна e_1^T , мы можем решать матричное уравнение

$$MH = T^T M \quad (49.2)$$

прямо относительно M и T при помощи метода, аналогичного методу § 43, где строки и столбцы поменялись ролями.

Мы все время старались подчеркнуть, что приведение к трехдиагональному виду потенциально неустойчиво, но мы не хотим слишком обескураживать читателя. Например, на АСЕ очень редко случалось, чтобы ведущие элементы достигали допуска 2^{-24} ($t = 48$). Если приведение H осуществлялось с двойной точностью, почти всегда было так, что ошибки, возникающие в этой части приведения, были значительно меньше ошибок, возникающих при приведении A к H с обычной точностью. Поэтому на практике комбинированная техника весьма эффективна.

Сравнение с методом Ланцоша

50. Метод Ланцоша § 35 позволяет нам переходить прямо от матрицы общего вида A к трехдиагональной матрице. Однако в этом методе мы должны вычислять обе последовательности c_i и c_i^* , и существенно осуществление переортогонализации. На каждой стадии существует опасность неустойчивости, и для безопасности мы все время вынуждены использовать арифметику с двойной точностью. Полное число умножений равно $4n^3$. Если на какой-либо стадии мы имеем слишком малые значения $c_i^T c_i^*$, мы вынуждены вернуться к началу, и снова есть, хотя и малая, вероятность второго срыва. В комбинированном методе § 49 нам нужно вернуться лишь к H . Таким образом, причин для использования комбинированной техники § 49 в предпочтение методу Ланцоша очень много. Во всех случаях, когда первый имеет какую-либо неустойчивость, в точности эта же неустойчивость появляется в несимметричном процессе Ланцоша.

Повторное исследование приведения к трехдиагональному виду

51. В § 44 мы заметили, что если на r -м шаге приведения нижней матрицы Хессенберга к трехдиагональному виду произойдет срыв, то мы должны вернуться к началу и вычислить $N_1 A N_1^{-1}$ с некоторой N_1 в надежде, что с новой матрицей срыва не произойдет. В применении к процессу Хессенберга, мы можем получить некоторые указания о выборе N_1 , т. е. о выборе c_1 .

Заметим, что первые $(2r + 1)$ элементов c_1 единственным образом определяют элементы $1, \dots, 2r + 2 - i$ векторов c_i ($i = 1, \dots, r - 1$). В случае $r = 3$ это иллюстрируется так:

$$\begin{array}{cccc}
 c_1 & c_2 & c_3 & c_4 \\
 \times & 0 & 0 & 0 \\
 \times & \times & 0 & 0 \\
 \times & \times & \times & 0 \\
 \times & \times & \times & \times \\
 \times & \times & \times & \times \\
 \times & \times & \times & \times \\
 \times & \times & \times & \times
 \end{array} \quad (51.1)$$

Следовательно, мы можем действовать таким образом. Выбираем первый элемент c_1 равным единице, а затем назначаем по два следующих элемента

за один раз. Каждый раз, как только мы добавляем два новых элемента, мы вычисляем возможные элементы c_i , как показано в (51.1). Если значительное взаимное уничтожение произойдет при вычислении $(r+1)$ -го элемента c_{r+1} , мы меняем последние два элемента c_1 и, следовательно, два последних вычисленных элемента каждого c_i ($i = 1, \dots, r+1$). Таким образом, мы можем избежать полного повторения вычислений, но только в том случае, если срыв не произошел после того, как мы достигли $c_{n/2}$. Имеется и другая слабость метода. Может случиться, что все элементы c_{r+1} малы в результате взаимного уничтожения. Если это так, то на самом деле не надо начинать сначала, но, к сожалению, на той стадии, когда $(r+1)$ -й элемент c_{r+1} вычислен, мы не знаем остальных элементов этого вектора.

Приведение матриц в верхней форме Хессенберга к форме Фробениуса

52. Наконец, рассмотрим приведение верхних матриц Хессенберга к форме Фробениуса (гл. 1, §§ 11–14). Обозначив поддиагональные элементы $h_{r+1, r}$ матрицы H через k_{r+1} , можно предположить, что $k_i \neq 0$ ($i = 2, \dots, n$).

Действительно, если $k_{r+1} = 0$, то

$$H = \begin{bmatrix} H_r & B_r \\ 0 & H_{n-r} \end{bmatrix}, \quad (52.1)$$

где H_r и H_{n-r} в верхней форме Хессенберга, и, следовательно,

$$\det(H - \lambda I) = \det(H_r - \lambda I) \det(H_{n-r} - \lambda I). \quad (52.2)$$

Заметим, что если $k_i \neq 0$ ($i = 2, \dots, n$), H не может быть неполной, так как ранг $(H - \lambda I)$ равен $(n-1)$ для каждого значения λ . С другой стороны, произвольное количество k_i могут быть равными нулю, даже если H полна. В последнем случае будем вычислять форму Фробениуса каждой соответствующей меньшей матрицы Хессенберга. Это во всех случаях лучше, так как дает нам факторизацию характеристического полинома.

Сначала допустим, что относительно k_i ничего, кроме того, что они отличны от нуля, не предполагается. Приведение состоит из $(n-1)$ основного шага. Перед началом r -го шага текущая матрица в случае $n = 6$, $r = 4$ имеет вид

$$r-1 \left\{ \begin{array}{ccc|ccc} 0 & 0 & 0 & \times & \times & \times \\ k_2 & 0 & 0 & \times & \times & \times \\ k_3 & 0 & & \times & \times & \times \\ \hline & & k_4 & \times & \times & \times \\ & & & k_5 & \times & \times \\ & & & & k_6 & \times \end{array} \right\}. \quad (52.3)$$

Поддиагональные элементы не меняются в течение процесса; r -й шаг заключается в следующем.

Для всех значений i от 1 до r выполнить шаги (i) и (ii):

(i) Вычисляем $n_{i, r+1} = h_{i, r}/k_{r+1}$.

(ii) Вычитаем из i -ой строки $n_{i, r+1} \times$ (строка $(r+1)$).

(iii) Для всех значений i от 1 до r прибавляем $k_{i+1}n_{i, r+1}$ к $h_{i, r+1}$.

Это завершает преобразование подобия.

Окончательная матрица X и ее диагональное преобразование подобия $F = D^{-1}XD$ имеют вид

$$X = \begin{bmatrix} 0 & 0 & 0 & \dots & x_0 \\ k_2 & 0 & 0 & \dots & x_1 \\ & k_3 & 0 & \dots & x_2 \\ \dots & \dots & \dots & \dots & \dots \\ & & & k_n & x_{n-1} \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & 0 & \dots & p_0 \\ 1 & 0 & 0 & \dots & p_1 \\ & 1 & 0 & \dots & p_2 \\ \dots & \dots & \dots & \dots & \dots \\ & & & 1 & p_{n-1} \end{bmatrix}, \quad (52.4)$$

где элементы D равны

$$1, k_2, k_2k_3, \dots, k_2k_3 \dots k_n, \quad (52.5)$$

а p_i равны

$$p_i = k_{i+2}k_{i+3} \dots k_n x_i. \quad (52.6)$$

Такое приведение требует приблизительно $\frac{1}{6} n^3$ умножений.

Влияние малых ведущих элементов

53. Если на некотором шаге $|k_{r+1}|$ значительно меньше $|h_{ir}|$, множители $n_{i, r+1}$ будут большими, и мы можем ожидать неустойчивости. К сожалению, мы не можем применять перестановки, так как они нарушили бы систему нулей. Следовательно, нет соответствующих преобразований, основанных на ортогональных матрицах. В самом деле, мы не можем, вообще говоря, получить каноническую форму Фробениуса при помощи ортогонального преобразования. На первый взгляд эффект малых ведущих элементов кажется значительно менее серьезным, чем в случае приведения к трехдиагональному виду. Действительно, предположим, что на r -м шаге k_{r+1} равен ε , а все остальные элементы порядка единицы. Мы видим, что в конце шага порядок величин, например, в случае $n = 5, r = 3$ будет таков:

$$\begin{bmatrix} 0 & 0 & 0 & \varepsilon^{-1} & \varepsilon^{-1} \\ 1 & 0 & 0 & \varepsilon^{-1} & \varepsilon^{-1} \\ & 1 & 0 & \varepsilon^{-1} & \varepsilon^{-1} \\ & & \varepsilon & 1 & 1 \\ & & & 1 & 1 \end{bmatrix}. \quad (53.1)$$

Здесь нет элементов порядка ε^{-2} . На самом деле даже элементы порядка ε^{-1} могут не быть вредными. Если мы выполним следующий шаг для матрицы (53.1), то окончательная матрица X и соответствующая F имеют следующие порядки величин элементов:

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & \varepsilon^{-1} \\ 1 & 0 & 0 & 0 & \varepsilon^{-1} \\ & 1 & 0 & 0 & \varepsilon^{-1} \\ & & \varepsilon & 0 & 1 \\ & & & 1 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ & 1 & 0 & 0 & 1 \\ & & 1 & 0 & 1 \\ & & & 1 & 1 \end{bmatrix}. \quad (53.2)$$

В F пропали все элементы порядка ε^{-1} !

Численный пример

54. Кажется, из обсуждения предыдущего параграфа, что использование малых ведущих элементов не вредно, в том смысле, что вычисленная форма Фробениуса может не сильно отличаться от матрицы, которая была бы получена при помощи точных вычислений. Рассмотрим простой пример, представленный в табл. 6.

Таблица 6

$$H = \begin{bmatrix} 10^0 (0,99995) & 10^{-5} (0,41325) \\ 10^{-5} (0,23125) & 10^1 (0,10001) \end{bmatrix}$$

$$X = \begin{bmatrix} 0 & -10^6 (0,43245) \\ 10^{-5} (0,23125) & 10^1 (0,20000) \end{bmatrix} \quad F = \begin{bmatrix} 0 & 10^1 (-0,10000) \\ 1 & 10^1 (0,20000) \end{bmatrix}$$

$$\text{Верная } F = \begin{bmatrix} 0 & 10^1 (-0,1000049 \dots) \\ 1 & 10^1 (0,200005) \end{bmatrix}$$

Мы видим, что вычисленная F — действительно правильна с рабочей точностью, несмотря на использование множителей порядка 10^5 при вычислении X . Тем не менее собственные значения вычисленной F не столь близки к собственным значениям точной F , которые, конечно, совпадают с собственными значениями H . Причина затруднения состоит в том, что матрица F значительно хуже обусловлена, чем H , и, следовательно, малые «ошибки» в элементах F ведут к весьма серьезным ошибкам в собственных значениях. Если мы оценим эквивалентное возмущение исходной матрицы H , вредный эффект преобразования сразу станет виден. В самом деле, если N — вычисленная матрица преобразования, то можно проверить, что

$$N^{-1}XN = \begin{bmatrix} 10^0 (0,999948125) & 10^2 (-0,1756873125) \\ 10^{-5} (0,23125) & 10^1 (0,1000051875) \end{bmatrix}. \quad (54.1)$$

Элемент (1, 2) имеет очень большое возмущение. В этом примере значительно лучше рассматривать $k_2 = h_{21}$ как нуль с рабочей точностью, так что собственные значения H получаются из двух меньших матриц Хессенберга 1-го порядка.

Общие замечания об устойчивости

55. Мы нашли, что при переходе к канонической форме Фробениуса часто имеет место значительное ухудшение чисел обусловленности. В самом деле все полные матрицы с одинаковой системой собственных значений, независимо от их обусловленности, имеют одинаковую форму Фробениуса. Рассмотрим, например, симметричную трехдиагональную матрицу W_{21}^- из гл. 5, § 45. Проблема собственных значений и собственных векторов для нее хорошо обусловлена, так как она симметрична, и ее собственные значения хорошо отделены. Тем не менее, как мы увидим в главе 7, соответствующая ей форма Фробениуса очень плохо обусловлена.

Мы описали приведение матрицы общего вида A к форме Фробениуса в две стадии.

Стадия 1. A приводится к верхней форме Хессенберга H при помощи устойчивых элементарных преобразований.

Стадия 2. H приводится к форме Фробениуса F при помощи неустойчивых элементарных преобразований.

Изучение устойчивости на этих двух стадиях очень различно. На практике мы обычно должны работать с большей точностью на второй стадии, чтобы ошибки, получающиеся в результате этой стадии, были столь же малы, как и на первой стадии.

В методе Данилевского (1937) эти две стадии скомбинированы. Этот метод состоит из $(n - 1)$ шага. Во время r -го вводятся нули в позициях $1, \dots, r, r + 2, \dots, n$ r -го столбца при помощи преобразования подобия с матрицей типа S_{r+1} (гл. 1, § 40). При помощи перестановок возможно обеспечить, чтобы величины $|s_{i, r+1}|$ ($i = r + 2, \dots, n$) были ограничены единицей, чего нельзя сделать с остальными $|s_{i, r+1}|$. При обычном описании метода Данилевского не обращают внимания на различия в численной устойчивости двух половин процесса.

Специальная верхняя форма Хессенберга

56. Приведение к форме Фробениуса, описанное в § 51, значительно упрощается, если исходная матрица Хессенберга имеет $k_i = 1$ ($i = 2, \dots, n$). В этом случае не требуется производить деления, а матрица X уже в требуемой форме F .

Если форма Хессенберга H матрицы общего вида A получается для того, чтобы затем получить форму Фробениуса, значительно удобнее сделать так, чтобы H имела единичные поддиагональные элементы, за исключением тех, которые необходимо равны нулю. Это может быть обеспечено в процессе Хессенберга § 29, если в качестве k_{r+1} брать единицу, кроме тех случаев, когда b_{r+1} равен нулю и мы *должны* брать k_{r+1} равным нулю.

Аналогично, если следовать § 11, нужно решать матричное уравнение

$$AN = NH \quad (56.1)$$

для H и N , однако теперь при предположении, что N — нижняя треугольная (первый столбец которой равен e_1), но уже не единичная нижняя треугольная и что H имеет единичные поддиагональные элементы. Так же, как в § 11, элементы N и H могут быть найдены столбец за столбцом. Каждый элемент N теперь задан в виде скалярного произведения, а каждый элемент h_{ij} матрицы H , кроме $h_{i+1, i}$, имеет вид скалярного произведения, деленного на n_{ii} . Мы имеем ту же экономию ошибок округления, и перестановки могут быть введены, как в § 12. Более того, присутствие единичных поддиагональных элементов значительно уменьшает число ошибок округления, которые делаются при дальнейшем приведении к форме Фробениуса.

Матрицы Хессенберга в настоящей формулировке могут быть получены из матриц §§ 11, 12 при помощи диагонального преобразования подобия. В то время как в предыдущих случаях элементы H были линейными по элементам A , поскольку элементы N — однородные функции нулевого порядка от a_{ij} , в настоящей формулировке h_{ij} степени $j - i + 1$. Это, очевидно, неудобно для вычислительной машины с фиксированной запятой. Предполагая, что мы используем плавающую запятую, найдем, что если A уравновешена, а $\|A\|$ сильно отличается от единицы, то H обычно будет совсем неуравновешенной. В то время как в формулировках §§ 11, 12 мы рассматривали поддиагональные элементы H как нули, если, например, они были меньше $2^{-t} \|A\|$, и разбивали H на две меньшие матрицы Хессенберга, мы не имеем подобного специфического критерия в настоящей формулировке.

Прямое определение характеристического полинома

57. Мы описали определение формы Фробениуса через преобразование подобия для последовательности изложения и для того, чтобы продемонстрировать связь с методом Данилевского. Однако, так как мы обычно используем более высокую точность при приведении к форме Фробениуса, чем при приведении к форме Хессенберга, матрицы, возникающие при приведении, не могут быть записаны в тех же ячейках, которые занимала матрица Хессенберга.

Более естественно определение самого характеристического полинома H . Этот полином может быть получен из рекуррентных соотношений, которыми последовательно определяются характеристические полиномы каждой последовательной ведущей подматрицы H_r ($r = 1, \dots, n$) матрицы H . Если обозначить $p_r(\lambda)$ характеристический полином подматрицы порядка r , то, разлагая $\det(H_r - \lambda I)$ по элементам r -го столбца, получим

$$\left. \begin{aligned} p_0(\lambda) &= 1, & p_1(\lambda) &= h_{11} - \lambda, \\ p_r(\lambda) &= (h_{rr} - \lambda)p_{r-1}(\lambda) - h_{r-1,r}k_r p_{r-2}(\lambda) + \\ &+ h_{r-2,r}k_r k_{r-1} p_{r-3}(\lambda) - \dots + (-1)^{r-1} h_{1r} k_r k_{r-1} \dots k_2 p_0(\lambda). \end{aligned} \right\} \quad (57.1)$$

Для определения коэффициентов $p_r(\lambda)$ нужно запомнить коэффициенты $p_i(\lambda)$ ($i = 1, \dots, r-1$). Для вычисления $p_r(\lambda)$ необходимо приблизительно $\frac{1}{2}r^2$ умножений, и следовательно, для полного вычисления харак-

теристического полинома H необходимо $\frac{1}{6}n^3$ умножений. Это столько же, как и при использовании преобразования подобия.

Если мы решим вычислить $p_n(\lambda)$ с двойной точностью, то и каждый из $p_r(\lambda)$ должен быть определен с двойной точностью. Однако, если предположить, что элементы H даны только с обычной точностью, то необходимы лишь умножения величин, заданных с двойной точностью, на величины, заданные с обычной точностью.

Никаких специальных трудностей не возникает, если некоторые из k_r малы или даже нули, хотя в последнем случае лучше иметь дело с соответствующими матрицами Хессенберга более низкого порядка. Снова процесс несколько упрощается, если k_r равны единице.

В общем, приведение к форме Фробениуса (или вычисление характеристического полинома) значительно менее удовлетворительно, чем приведение к трехдиагональному виду. Для многих выглядящих весьма безобидно расположений собственных значений форма Фробениуса очень плохо обусловлена, и довольно часто бывает недостаточно даже двойной точности вычислений. Мы более полно опишем этот вопрос в главе 7.

Дополнительные замечания

Несмотря на то, что правило выбора ведущего элемента было широко обсуждено в связи с методом Гаусса, его роль в преобразованиях подобия обсуждалась сравнительно мало; на самом же деле здесь это даже более важно. Если A имеет линейные элементарные делители, то мы имеем $AX = X\Lambda$ при некоторой X , и следовательно,

$$(D_1 A D_1^{-1}) D_1 X D_2 = D_1 X D_2 \Lambda$$

с любыми диагональными матрицами D_1 и D_2 . (Матрица D_2 дает только нормировку собственных векторов.) Мы хотим подобрать D_1 так, чтобы $k(D_1 X D_2)$ было как можно меньше; эта проблема соответствует уравниванию в случае обращения матрицы.

Бауэр (1963) дал глубокое обсуждение этой проблемы. На практике трудность состоит в том, что необходимо выбрать D_1 перед тем, как X определена. Поэтому проблема в некотором отношении более трудная, чем проблема уравнивания A для обращения матрицы. Осборн (1960) обсуждал эту проблему с более эмпирической точки зрения. Снова, как и в случае проблемы обращения матрицы, нужно подчеркнуть, что необходимость уравнивания не связана с использованием исключения с главным элементом. Его роль может быть столь же велика при использовании ортогональных преобразований.

Несмотря на то, что компактные схемы треугольной факторизации использовались давно, прямое приведение к форме Хессенберга, обсуждавшееся в §§ 11 и 12, выглядит новым. Оно впервые использовалось на АСЕ в 1959 г. (включая применение перестановок); вычисления производились с использованием арифметики с двойной точностью и учетверенным накоплением скалярных произведений.

Обсуждение формальных аспектов обобщенных процессов Хессенберга в §§ 26—35 основано на работе Хаусхолдера и Бауэра (1959). Метод Ланцоша (1950) обсуждался в большом количестве работ, в частности в статьях Брукера и Сумнера (1956), Козей и Грегори (1961), Грегори (1958), Россера и др. (1951), Рутисхаузера (1953), Уилкинсона (1958b). Из этих последовательных подходов постепенно возникла надежная автоматическая процедура для определения собственных значений и собственных векторов, и история этого алгоритма особенно поучительна для численного анализа. Ирония заключается в том, что к тому времени, когда практический процесс был полностью понят, он был вытеснен методом § 43.

Общее обсуждение методов для приведения матрицы общего вида к трехдиагональному виду было дано Бауэром (1959), и методы, описанные в этой работе, могут привести к алгоритмам, лучшим, чем тот, который мы описали.

СОБСТВЕННЫЕ ЗНАЧЕНИЯ МАТРИЦ СПЕЦИАЛЬНОГО ВИДА

Введение

1. В этой главе рассматривается решение проблемы собственных значений для матриц специального вида, рассмотренных в предыдущей главе. Пусть A обозначает матрицу в одной из этих форм. Нам нужно определить нули функции

$$f(z) = \det(A - zE). \quad (1.1)$$

Так как эта функция является полиномом, то все соответствующие методы должны быть существенно итерационными по характеру.

Методы, обсуждающиеся в этой главе, зависят существенно от вычисления значений $f(z)$ и, возможно, ее производных, для последовательностей чисел, стремящихся к ее корням. Различные методы отличаются:

- (i) алгоритмами для вычисления значений $f(z)$ и ее производных,
- (ii) процессами получения членов последовательности,
- (iii) приемами, которые используются для того, чтобы избежать сходимости к одному и тому же собственному значению более одного раза.

Здесь (i) существенно зависит от формы матрицы A , а (ii) и (iii) почти не зависят от нее.

Во всех методах окончательное расположение нулей зависит от вычисленных значений, полученных в окрестности этих нулей. Очевидно, что фундаментальную роль играет точность, с которой мы можем вычислить $f(z)$ в этих окрестностях. Поэтому мы сначала рассмотрим методы вычисления значений $f(z)$.

Явно заданная полиномиальная форма

2. Самой простой из форм, рассмотренных в главе 6, является форма Фробениуса. Она дает нам либо характеристический полином в чистом виде, либо его факторизацию в два или более полиномов. Коэффициент при старшей степени z равен единице, так что мы рассмотрим вычисления $f(z)$ вида

$$f(z) = z^n + a_{n-1}z^{n-1} + \dots + a_0. \quad (2.1)$$

Простейшим методом вычисления является метод, обычно известный под названием схемы Горнера. Для значения $z = \alpha$ вычисляем $s_r(\alpha)$ ($r = n, \dots, 0$)

$$s_n(\alpha) = 1, \quad s_r(\alpha) = \alpha s_{r+1}(\alpha) + a_r \quad (r = n-1, \dots, 0). \quad (2.2)$$

Очевидно, что $f(\alpha) = s_0(\alpha)$. Так как $s_0(\alpha)$ удовлетворяет уравнению

$$s_0(\alpha) = \prod_{i=1}^n (\alpha - \lambda_i), \quad (2.3)$$

требуемые значения могут находиться в очень широкой области, и вычисления с плавающей запятой почти необходимы.

Вычисленные значения удовлетворяют соотношениям

$$\left. \begin{aligned} s_n(\alpha) &= 1, \\ s_r(\alpha) &= fl[\alpha s_{r+1}(\alpha) + a_r] \equiv \\ &\equiv \alpha s_{r+1}(\alpha)(1 + \varepsilon_1) + a_r(1 + \varepsilon_2) \quad (r = n-1, \dots, 0), \end{aligned} \right\} \quad (2.4)$$

где

$$|\varepsilon_1| < 2 \cdot 2^{-t_1}, \quad |\varepsilon_2| < 2^{-t_1}. \quad (2.5)$$

Объединяя эти результаты, имеем

$$s_0(\alpha) = (1 + E_n) \alpha^n + a_{n-1}(1 + E_{n-1}) \alpha^{n-1} + \dots + \alpha_0(1 + E_0), \quad (2.6)$$

причем, очевидно,

$$|E_r| < (2r + 1) 2^{-t_1}. \quad (2.7)$$

Вычисленное значение $s_0(\alpha)$ поэтому является точным значением полинома с коэффициентами $a_r(1 + E_r)$. Следовательно, несмотря на то, что при вычислении $s_0(\alpha)$, конечно, могут иметь место взаимные уничтожения, ошибки могут быть отнесены к сравнительно малым возмущениям коэффициентов.

Заметим, однако, что полином с коэффициентами $a_r(1 + E_r)$ не даст те же самые значения $s_r(\alpha)$ ($r = 1, \dots, n-1$). В самом деле из (2.4) имеем

$$s_r(\alpha) \equiv \alpha s_{r+1}(\alpha)(1 + \varepsilon_1) + a_r(1 + \varepsilon_2) \equiv \alpha s_{r+1}(\alpha) + [a_r(1 + \varepsilon_2) + \alpha s_{r+1}(\alpha) \varepsilon_1], \quad (2.8)$$

что показывает, что полином, дающий все вычисленные $s_r(\alpha)$, имеет коэффициенты, равные

$$a'_r = a_r(1 + \varepsilon_2) + \alpha s_{r+1}(\alpha) \varepsilon_1. \quad (2.9)$$

Если изменение от a_r к a'_r приводит к малому относительному возмущению, то $\alpha s_{r+1}(\alpha)$ будет того же порядка, что и a_r , или меньше при всех r . Для многих полиномов это далеко не так. Однако если все нули λ_i ($i = 1, \dots, n$) вещественные и имеют один знак, эти условия всегда выполнены для α в окрестности любого нуля. Чтобы доказать это, заметим, что

$$f(z) \equiv (z - \alpha) \sum_1^n s_r(\alpha) z^{r-1} + s_0(\alpha). \quad (2.10)$$

Из (2.1) имеем

$$a_r = \pm \sum \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_{n-r}}, \quad (2.11)$$

откуда, положив $\alpha = \lambda_k$ в (2.10) и замечая, что $s_0(\lambda_k) = 0$, получим

$$s_r(\lambda_k) = \pm \sum \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_{n-r}} \quad (\lambda_{i_j} \neq \lambda_k). \quad (2.12)$$

Отсюда видно, что $s_r(\lambda_k)$ того же знака, что и a_r , и не больше его по абсолютной величине. В силу (2.2) это означает, что $|\lambda_k s_{r+1}(\lambda_k)| \leq |a_r|$. Следовательно, из (2.9) имеем

$$|a'_r - a_r| \leq |a_r \varepsilon_2| + |\lambda_k s_{r+1}(\lambda_k) \varepsilon_1| \leq |a_r| 2^{-t_1} + 2 |a_r| 2^{-t_1} = 3 |a_r| 2^{-t_1}. \quad (2.13)$$

Вообще говоря, при вычислении значений полиномов трудно получить выигрыш от накопления скалярных произведений. Если мы хотим избегать множителей, имеющих постоянно растущее число значащих цифр, мы должны округлять каждое $s_r(\alpha)$ перед вычислением следующего. Имеем

$$s_r(\alpha) = fl_2[\alpha s_{r+1}(\alpha) + a_r] \equiv [\alpha s_{r+1}(\alpha)(1 + \varepsilon_1) + a_r(1 + \varepsilon_2)](1 + \varepsilon), \quad (2.14)$$

где

$$|\varepsilon_1| < 3 \cdot 2^{-2t_2}, \quad |\varepsilon_2| < (1,5) 2^{-2t_2}, \quad |\varepsilon| \leq 2^{-t}. \quad (2.15)$$

Следовательно, эквивалентное возмущение a_r будет порядка 2^{-t} , если только не будет

$$|s_r(\alpha)| \ll |a_r|, \quad (2.16)$$

т. е. если не происходит большого взаимного уничтожения при вычислении s_r .

Значения производных $f(z)$ можно вычислять при помощи рекуррентных соотношений, подобных (2.2). В самом деле, продифференцировав эти уравнения, получаем

$$s'_r(\alpha) = \alpha s'_{r+1}(\alpha) + s_{r+1}(\alpha), \quad s'_r = \alpha s'_{r+1}(\alpha) + 2s'_{r+1}(\alpha), \quad (2.17)$$

и можно вычислять первую и вторую производную вместе с самой функцией.

3. Мы показали, что во всех случаях вычисленные значения $s_0(\alpha)$ функции $f(\alpha)$ удовлетворяют соотношению

$$s_0(\alpha) = \sum a_r(1 + E_r)\alpha^r, \quad |E_r| < (2r + 1)2^{-t_1}, \quad (3.1)$$

и указали, что для некоторых определенных расположений нулей верхняя граница для $|E_r|$ может быть значительно меньше. Во всяком случае, принимая во внимание статистический эффект, мы можем ожидать, что E_r должна удовлетворять оценке типа

$$|E_r| < (2r + 1)^{1/2} 2^{-t_1}. \quad (3.2)$$

Если явный полином был получен при помощи приведения матрицы общего вида к форме Фробениуса, то ошибки в коэффициентах, возникающие из-за этого приведения, часто будут значительно больше ошибок, соответствующих (3.2), и фактическое вычисление значений полинома вряд ли будет слабым звеном цепи.

Заметим, что возмущения E_r являются функциями α . Если мы обозначим нули $\sum a_r(1 + E_r)z^r$ через $\lambda_i^{(\alpha)}$ ($i = 1, 2, \dots, n$), то

$$s_0(\alpha) = \prod (\alpha - \lambda_i^{(\alpha)}), \quad (3.3)$$

в то время как верное значение $f(\alpha)$ равно

$$f(\alpha) = \prod (\alpha - \lambda_i). \quad (3.4)$$

Числа обусловленности явно заданных полиномов

4. Слабость методов определения собственных значений, основанных на приведении к форме Фробениуса, почти целиком является следствием того, что нули многих полиномов чрезвычайно чувствительны к малым относительным возмущениям коэффициентов. Это верно не только для поли-

номов, имеющих кратные или патологически близкие нули, но также для многих полиномов, расположение нулей которых на первый взгляд кажется очень хорошим. Это обсуждалось Уилкинсоном (1959а), а также составило основное содержание главы 2 работы Уилкинсона (1963b). Здесь мы ограничимся кратким суммированием результатов.

Рассмотрим нули полинома

$$g(z) = f(z) + \delta a_k z^k = \sum a_r z^r + \delta a_k z^k. \quad (4.1)$$

Верны следующие простые результаты:

(i) Если λ_i — простой корень $f(z)$, то соответствующий корень $(\lambda_i + \delta\lambda_i)$ полинома $g(z)$ удовлетворяет соотношению

$$\delta\lambda_i \sim -\delta a_k \lambda_i^{k/f'(\lambda_i)} = -\delta a_k \lambda_i^{k/\prod_{j \neq i} (\lambda_i - \lambda_j)}. \quad (4.2)$$

(ii) Если λ_i — корень $f(z)$ кратности m , то существуют m соответствующих корней $(\lambda_i + \delta\lambda_i)$ полинома $g(z)$, где $\delta\lambda_i$ это m величин, удовлетворяющих соотношениям

$$\delta\lambda_i \sim [-m! \delta a_k \lambda_i^{k/f^{(m)}(\lambda_i)}]^{1/m} = [-\delta a_k \lambda_i^{k/\prod_{j \neq i} (\lambda_i - \lambda_j)}]^{1/m}. \quad (4.3)$$

Итак, в случае простого корня $|\lambda_i^{k/\prod_{j \neq i} (\lambda_i - \lambda_j)}|$ является числом обусловленности для корня λ_i . Этот результат можно сравнить с общей теорией возмущений для матриц в форме Фробениуса. В гл. 1, § 11 мы видели, что правый собственный вектор x_i , соответствующий одному из вариантов формы Фробениуса, имеет вид

$$x_i^T = (\lambda_i^{n-1}, \lambda_i^{n-2}, \dots, 1). \quad (4.4)$$

Подставляя a_r на место p_r из главы 1, получим соответствующий левый собственный вектор y_i в виде

$$y_i^T = (1; \lambda_i + a_{n-1}; \lambda_i^2 + a_{n-1}\lambda_i + a_{n-2}; \dots; \lambda_i^{n-1} + a_{n-1}\lambda_i^{n-2} + \dots + a_1). \quad (4.5)$$

Очевидно, что мы имеем $y_i^T x_i = f'(\lambda_i)$, и, следовательно, изменение $\delta\lambda_i$ собственного значения λ_i , соответствующее возмущению εB матрицы Фробениуса, удовлетворяет соотношению

$$\delta\lambda_i \sim y_i^T (\varepsilon B) x_i / y_i^T x_i. \quad (4.6)$$

Если взять за εB матрицу, соответствующую одному возмущению δa_k элемента a_k , то соотношение (4.6) немедленно даст (4.2).

Несколько типичных расположений корней

5. Соотношение (4.3) показывает, что кратные корни очень чувствительны к малым возмущениям коэффициентов, и это хорошо известно. Гораздо менее широко известно, что чрезвычайная чувствительность часто связана с совершенно безобидными на первый взгляд расположениями корней. Рассмотрим поэтому эту проблему для всех расположений, которые рассматривались в гл. 6, § 23, в связи с методом Крылова. Для иллюстрации будем рассматривать влияние возмущения $2^{-t} a_k$ коэффициента a_k .

(i) Линейное расположение $\lambda_r = 21 - r$. Имеем

$$|\delta\lambda_i| \sim 2^{-i} |a_k| (21 - i)^k / (20 - i)! (i - 1)!, \quad (5.4)$$

и порядки величин коэффициентов видны из (5.2):

$$\begin{aligned} x^{20} + 10^3 x^{19} + 10^5 x^{18} + 10^7 x^{17} + 10^8 x^{16} + 10^{10} x^{15} + 10^{11} x^{14} + \\ + 10^{12} x^{13} + 10^{13} x^{12} + 10^{15} x^{11} + 10^{16} x^{10} + 10^{17} x^9 + 10^{17} x^8 + 10^{18} x^7 + \\ + 10^{19} x^6 + 10^{19} x^5 + 10^{19} x^4 + 10^{20} x^3 + 10^{20} x^2 + 10^{19} x + 10^{19}. \end{aligned} \quad (5.2)$$

Тривиальное, но скучное исследование показывает, что наиболее чувствителен корень $x = 16$ ($i = 5$) по отношению к возмущению a_{15} . Имеем

$$|\delta\lambda_5| \sim 2^{-i} 10^{10} (16)^{15} / 15! 4! \approx 2^{-i} (0,38) 10^{15}, \quad (5.3)$$

что показывает, что пока 2^{-i} не будет значительно меньше, чем 10^{-14} , возмущенный корень будет иметь малое отношение к исходному. Большинство наибольших корней очень чувствительны к возмущениям старших коэффициентов.

Т а б л и ц а 1

Нули полинома $(x-1)(x-2)\dots(x-20) - 2^{-23}x^{19}$,
верные с 9 десятичными знаками

1,00000 0000	6,00000 6944	10,09526 6145 $\pm 0,64350$ 0904i
2,00000 0000	6,99969 7234	11,79363 3881 $\pm 1,65232$ 9728i
3,00000 0000	8,00726 7603	13,99235 8137 $\pm 2,51883$ 0070i
4,00000 0000	8,91725 0249	16,73073 7466 $\pm 2,81262$ 4894i
4,99999 9928	20,84690 8101	19,50243 9400 $\pm 1,94033$ 0347i

Это иллюстрируется в табл. 1, где показан эффект одного возмущения 2^{-23} в коэффициенте при x^{19} , что составляет приблизительно 2^{-30} его величины. Десять корней стали комплексными и имеют весьма значительную мнимую часть. С другой стороны, малые корни не чувствительны, что видно из (5.1).

(ii) Линейное расположение $\lambda_r = 11 - r$. Имеем

$$|\delta\lambda_i| \sim 2^{-i} |a_k| (11 - i)^k / (20 - i)! (i - 1)! \quad (5.4)$$

и сравнение с (5.1) показывает, что множитель $(21 - i)^k$ заменен на $(11 - i)^k$. Порядки величин коэффициентов показаны в (5.5):

$$\begin{aligned} x^{20} + 10x^{19} + 10^3 x^{18} + 10^4 x^{17} + 10^5 x^{16} + 10^6 x^{15} + 10^7 x^{14} + \\ + 10^8 x^{13} + 10^9 x^{12} + 10^9 x^{11} + 10^{10} x^{10} + 10^{11} x^9 + 10^{11} x^8 + \\ + 10^{12} x^7 + 10^{12} x^6 + 10^{12} x^5 + 10^{12} x^4 + 10^{12} x^3 + 10^{12} x^2 + 10^{12} x. \end{aligned} \quad (5.5)$$

Наибольшей чувствительностью обладает λ_3 по отношению к возмущениям a_{16} . Имеем

$$|\delta\lambda_3| \sim 2^{-i} |a_{16}| 8^{16} / 2! 17! \approx 2^{-i} (0,4) 10^5. \quad (5.6)$$

Поэтому наш полином значительно лучше обусловлен, чем соответствующий полином из (i). Заметим, что если A имеет собственные значения $(21 - i)$, то $(A - 10I)$ имеет собственные значения $(11 - i)$. Следовательно, сравнительно тривиальное изменение исходной матрицы имеет большое влияние на обусловленность соответствующей формы Фробениуса. Это не так при приведении к трехдиагональному виду и к форме Хессенберга, описанным в главе 6; для них изменение на kI исходной матрицы приводит к изменению на kI в приведенной матрице, и следовательно,

обусловленность приведенной матрицы не зависит от преобразований такого рода.

(iii) Геометрическое расположение $\lambda_r = 2^{-r}$. Коэффициенты полинома очень сильно различаются по величине; их порядки указаны в (5.7):

$$\begin{aligned} x^{20} + x^{19} + 2^{-1}x^{18} + 2^{-4}x^{17} + 2^{-8}x^{16} + 2^{-13}x^{15} + \\ + 2^{-19}x^{14} + 2^{-26}x^{13} + 2^{-34}x^{12} + 2^{-43}x^{11} + 2^{-53}x^{10} + \\ + 2^{-64}x^9 + 2^{-76}x^8 + 2^{-89}x^7 + 2^{-103}x^6 + 2^{-118}x^5 + \\ + 2^{-134}x^4 + 2^{-151}x^3 + 2^{-169}x^2 + 2^{-189}x + 2^{-209}. \end{aligned} \quad (5.7)$$

При двух предыдущих расположениях мы рассматривали абсолютные изменения корней. Сейчас более естественно рассматривать относительные возмущения.

Имеем

$$|\delta\lambda_i/\lambda_i| \sim 2^{-t} |a_k| 2^{(1-k)i} / |PQ|, \quad (5.8)$$

где

$$P = (2^{-i} - 2^{-1})(2^{-i} - 2^{-2}) \dots (2^{-i} - 2^{-i+1}), \quad (5.9)$$

$$Q = (2^{-i} - 2^{-i-1})(2^{-i} - 2^{-i-2}) \dots (2^{-i} - 2^{-20}). \quad (5.10)$$

Тогда

$$|P| = 2^{-\frac{1}{2}i(i-1)} [(1 - 2^{-1})(1 - 2^{-2}) \dots (1 - 2^{-i+1})], \quad (5.11)$$

$$|Q| = 2^{-i(20-i)} [(1 - 2^{-1})(1 - 2^{-2}) \dots (1 - 2^{-20+i})], \quad (5.12)$$

и оба выражения в квадратных скобках стремятся к бесконечному произведению

$$(1 - 2^{-1})(1 - 2^{-2})(1 - 2^{-3}) \dots \quad (5.13)$$

Довольно грубая оценка показывает, что значение этого произведения лежит между 1/2 и 1/4, и, следовательно, имеем

$$|\delta\lambda_i/\lambda_i| < 2^{-t} |a_k| 2^{\theta+4}, \quad \text{где} \quad \theta = \frac{1}{2}i(41 - 2k - i). \quad (5.14)$$

При фиксированном значении k максимум достигается при $i = 20 - k$, так что имеем

$$|\delta\lambda_i/\lambda_i| < 2^{-t} |a_k| 2^{4 + \frac{1}{2}(20-k)(21-k)}, \quad (5.15)$$

и из (5.7) можно проверить, что

$$|a_k| \leq 4 \cdot 2^{-\frac{1}{2}(20-k)(21-k)}. \quad (5.16)$$

Следовательно, для относительного возмущения собственного значения, вызванного относительным возмущением 2^{-t} любого коэффициента, окончательно имеем

$$\left| \frac{\delta\lambda_i}{\lambda_i} \right| < 2^{6-t}. \quad (5.17)$$

Поэтому корни довольно хорошо обусловлены по отношению к таким возмущениям, которые сравнимы с ошибками, сделанными при вычислении значений полинома.

Естественно спросить, будем ли мы обычно получать коэффициенты характеристического уравнения с малой относительной ошибкой, если собственные значения широко меняются по величине. К сожалению, ответом, вообще говоря, будет «нет». В качестве очень простого примера рассмотрим матрицу

$$A = \begin{bmatrix} 0,63521 & 0,81356 \\ 0,59592 & 0,76325 \end{bmatrix}, \quad (5.18)$$

собственные значения которой с пятью точными значащими цифрами равны: $\lambda_1 = 10^1 (0,13985)$, $\lambda_2 = 10^{-5} (0,52610)$. Мы можем привести ее к форме Фробениуса при помощи преобразования подобия $M_1^{-1} A M_1$, где M_1 равна

$$M_1 = \begin{bmatrix} 1 & 10^1 (-0,12808) \\ 0 & 1 \end{bmatrix}, \quad (5.19)$$

но, несмотря на использование вычислений с плавающей запятой, вычисленная преобразованная матрица равна

$$\begin{bmatrix} 10^1 (0,13985) & 10^{-4} (-0,2000) \\ 0,59592 & 0 \end{bmatrix}. \quad (5.20)$$

При вычислении элемента в позиции (1, 2) происходит значительное взаимное уничтожение; вычисленный элемент, дающий свободный член в характеристическом уравнении, имеет только одну правильную значащую цифру. Вообще, если имеются большие различия в величинах собственных значений, элементы в форме Фробениуса, как правило, получаются с абсолютными ошибками, порядок которых почти одинаков для всех элементов, и это весьма фатально для относительной точности малых собственных значений. Например, очевидно, что абсолютная ошибка порядка 10^{-10} в свободном члене полинома, соответствующего расположению (iii), меняет произведение его корней от 2^{-210} до $(2^{-210} + 10^{-10})$, т. е. приблизительно в 10^{33} раз.

Для специальных типов матриц такое ухудшение не имеет места. Простым примером является матрица

$$\begin{bmatrix} 10^0 (0,2632) & 10^{-5} (0,3125) \\ 10^0 (0,4315) & 10^{-5} (0,2873) \end{bmatrix}. \quad (5.21)$$

Несмотря на то, что одно ее собственное значение порядка единицы, а другое порядка 10^{-5} , при приведении к форме Фробениуса не происходит взаимного уничтожения.

(iv) Расположение по окружности $\lambda_r = \exp(r\pi i/20)$. Соответствующий полином равен $z^{20} - 1$, и мы имеем

$$|\delta\lambda_i| \sim |\delta a_k| |\lambda_i|^k / |f'(\lambda_i)| = \frac{1}{20} |\delta a_k|. \quad (5.22)$$

Нули этого полинома очень хорошо обусловлены. (Так как почти все коэффициенты равны нулю, мы в этом случае не можем рассматривать относительное возмущение.)

Окончательная оценка метода Крылова

6. Анализ предыдущего параграфа завершает оценку, сделанную в гл. 6, § 25, о степени эффективности метода Крылова в применении к матрицам с различными расположениями собственных значений, приведенными в § 5. При расположении (i) использование 62 двоичных знаков в нормализованном методе Крылова абсолютно недостаточно. Нули вычисленного характеристического полинома не имеют никакого отношения к собственным значениям исходной матрицы. Несмотря на это, для тесно связанного расположения (ii) вычисленный характеристический полином значительно более точен и, более того, значительно лучше обусловлен.

Мы видели в гл. 6, § 25, что для расположения типа (iii) вычисления с 62 двоичными знаками дают совсем неузнаваемый вычисленный характеристический полином. Тот факт, что нули верного характеристического уравнения сравнительно нечувствительны к малым относительным изменениям коэффициентов, не имеет отношения к делу. Для таких расположений метод Крылова непрактичен.

Рассматривая расположение (iv), находим, что не только метод Крылова дает очень точное определение характеристического полинома, но и что характеристическое уравнение очень хорошо обусловлено. Метод Крылова, даже с одинарной точностью, очень эффективен для матриц с собственными значениями этого типа.

Суммируя, мы можем сказать, что, несмотря на то, что метод Крылова эффективен для матриц, собственные значения которых «хорошо распределены» в комплексной плоскости, он имеет серьезные ограничения для использования в качестве метода широкого назначения.

Общие замечания о явно заданных полиномах

7. Очевидно, что для расположения типа $\lambda_i = 21 - i$ ($i = 1, \dots, 20$) необходимо вычислять форму Фробениуса, используя арифметику с высокой точностью, если мы хотим, чтобы ее собственные значения были близки к собственным значениям исходной матрицы, из которой она получена. Это верно независимо от использованного метода. Далее, так как ошибки при вычислении $f(\alpha)$ соответствуют возмущениям коэффициентов, определяемым (2.6) и (2.7), вычисление $f(z)$ нужно производить с той же высокой точностью.

Интересно, например, заметить, что вычисленные значения $f(z)$ для расположения (i) § 5 обычно не будут иметь ни одной правильной значащей цифры при значениях z в окрестности $z = 20$, если только не используется арифметика с высокой точностью. Для иллюстрации рассмотрим вычисление $f(z)$ при $z = 20,00345645$ с использованием 10 десятичных знаков в арифметике с плавающей запятой. На втором шаге схемы Горнера вычисляем

$$s_{19}(z) = zs_{20}(z) + a_{19} = 20,00345645 \times 1 - 210,0. \quad (7.1)$$

Точное значение равно $-189,99654355$, но если мы используем только 10 цифр, его надо округлить до $-189,9965436$. Даже если больше никаких других ошибок не будет сделано в дальнейшем, окончательная

ошибка, вызванная этим округлением, будет равна

$$5 \times 10^{-8} (20,00345645)^{19} \approx (2,6) 10^{17}. \quad (7.2)$$

Если мы для краткости напомним $x = 0,00345645$, точное значение $f(z)$ будет равно

$$x(x+1)(2+x)\dots(19+x), \quad (7.3)$$

и оно лежит между $19! x$ и $20! x$, т. е. между $(4,2) 10^{14}$ и $(8,4) 10^{15}$. Наша первая ошибка округления меняет вычисленное значение на величину, значительно превосходящую точное значение!

В окрестности плохо обусловленных корней вычисленные значения $f(z)$, соответствующие монотонной последовательности z , могут быть совсем немонотонными. Это иллюстрируется в табл. 2, где приводятся

		Таблица 2
z	Вычисленное $f(z)$	Точное $f(z)$
$10 + 2^{-42}$	$+0,63811\dots$	$10^{-6} (+0,16501\dots)$
$10 + 2 \times 2^{-42}$	$+0,57126\dots$	$10^{-6} (+0,33003\dots)$
$10 + 3 \times 2^{-42}$	$-0,31649\dots$	$10^{-6} (+0,49505\dots)$
$10 + 2^{-28} + 7 \times 2^{-42}$	$+0,29389\dots$	$10^{-2} (+0,27048\dots)$
$10 + 2^{-23} + 7 \times 2^{-42}$	$+0,70396\dots$	$10^{-1} (+0,86518\dots)$
$10 + 2^{-18} + 7 \times 2^{-42}$	$10^1 (+0,33456\dots)$	$10^1 (+0,27685\dots)$
$10 + 2^{-13} + 7 \times 2^{-42}$	$10^2 (+0,89316\dots)$	$10^2 (+0,88608\dots)$

вычисленные значения полинома $\prod_{r=1}^{12} (z - r)$ для значений z в окрестности 10. Включение члена $7 \cdot 2^{-42}$ в последние четыре значения z было необходимо для того, чтобы ошибки округления были типичными, ибо иначе z оканчивалось бы большим количеством нулей. Значения полинома вычислялись по его точному явному представлению с использованием арифметики с плавающей запятой с мантиссой в 46 двоичных знаков.

(Мы не использовали $\prod_{r=1}^{20} (z - r)$, так как при выбранном числе значащих цифр этот полином не может быть записан точно.)

Для значений z от 10 и приблизительно до $10 + 2^{-20}$ все вычисленные значения порядка единицы по абсолютной величине и никак не связаны с точными значениями. Это происходит потому, что вычисленные значения полностью определяются ошибками округления, которые, как мы видели в (2.13), эквивалентны возмущениям коэффициентов a_r порядка вплоть до $3 |a_r| 2^{-l_1}$. Для значений z , более удаленных от 10, мы получаем несколько правильных значащих цифр. Действительно, если $z = 10 + x$ и $|x|$ порядка 2^{-k} , то, вообще говоря, вычисленное значение $f(z)$ имеет приблизительно $20 - k$ правильных значащих цифр, хотя, конечно, исключительно благоприятное округление может дать значительно лучшее значение. Трудно ожидать, что какой-либо итерационный метод, основанный на значениях, вычисленных по явно заданному полиному с использованием 46-значной мантиссы, выделит нуль при $z = 10$ с ошибкой, меньшей 2^{-20} .

Трехдиагональные матрицы

8. Возвратимся к вычислению $\det(A - zI)$, где A — трехдиагональная матрица. Обозначим

$$a_{ii} = \alpha_i, \quad a_{i, i+1} = \beta_{i+1}, \quad a_{i+1, i} = \gamma_{i+1}, \quad (8.1)$$

и пусть $p_r(z)$ — главные ведущие миноры $(A - zI)$ порядка r . Раскладывая $p_r(z)$ по его последней строке, получим

$$p_r(z) = (\alpha_r - z)p_{r-1}(z) - \beta_r \gamma_r p_{r-2}(z) \quad (r = 2, \dots, n), \quad (8.2)$$

где

$$p_0(z) = 1, \quad p_1(z) = \alpha_1 - z. \quad (8.3)$$

Произведения $\beta_r \gamma_r$ могут быть вычислены раз и навсегда. Можно предполагать, что $\beta_r \gamma_r \neq 0$, так как иначе мы можем иметь дело с трехдиагональными матрицами более низкого порядка. Для каждого вычисления необходимо $2n$ умножений.

Если требуются значения производных $\det(A - zI)$, то они могут быть вычислены одновременно с вычислением значений самого определителя. В самом деле, дифференцируя (8.2), имеем

$$p'_r(z) = (\alpha_r - z)p'_{r-1}(z) - \beta_r \gamma_r p'_{r-2}(z) - p_{r-1}(z), \quad (8.4)$$

и

$$p''_r(z) = (\alpha_r - z)p''_{r-1}(z) - \beta_r \gamma_r p''_{r-2}(z) - 2p'_{r-1}(z). \quad (8.5)$$

Так как точность вычисленных корней определяется по существу точностью вычисленных значений самой функции, мы ограничимся анализом ошибок алгоритма, описанного в (8.2). На практике почти необходимо использование арифметики с плавающей запятой. Вычисленные величины удовлетворяют соотношениям

$$\begin{aligned} p_r(z) &= fl[(\alpha_r - z)p_{r-1}(z) - \beta_r \gamma_r p_{r-2}(z)] \equiv \\ &\equiv (\alpha_r - z)(1 + \varepsilon_1)p_{r-1}(z) - \beta_r \gamma_r (1 + \varepsilon_2)p_{r-2}(z), \end{aligned} \quad (8.6)$$

где

$$|\varepsilon_1| < 3 \cdot 2^{-t_1}, \quad |\varepsilon_2| < 3 \cdot 2^{-t_1}. \quad (8.7)$$

Следовательно, как в симметричном случае (гл. 5, § 40), вычисленная последовательность является точной для трехдиагональной матрицы с элементами α'_r , β'_r , γ'_r , где

$$\alpha'_r - z = (\alpha_r - z)(1 + \varepsilon_1), \quad \beta'_r \gamma'_r = \beta_r \gamma_r (1 + \varepsilon_2). \quad (8.8)$$

Поэтому можно предположить, что

$$\beta'_r = \beta_r (1 + \varepsilon_2)^{1/2}, \quad \gamma'_r = \gamma_r (1 + \varepsilon_2)^{1/2}, \quad (8.9)$$

так что β'_r и γ'_r соответствуют малым относительным ошибкам. С другой стороны, имеем

$$\alpha'_r - \alpha_r = (\alpha_r - z)\varepsilon_1, \quad (8.10)$$

и мы не можем гарантировать, что это соответствует малым относительным ошибкам в α_r . Однако, мы можем ограничиться лишь такими значениями z , что

$$|z| \leq \|A\|_\infty = \max(|\gamma_i| + |\alpha_i| + |\beta_{i+1}|), \quad (8.11)$$

так что возмущения α_r малы по сравнению с нормой A .

9. Заметим, что если $\tilde{A}$ — какая-либо матрица, полученная при помощи диагонального преобразования подобия матрицы A , то α_r и произведения $\beta_r \gamma_r$ для $\tilde{A}$ и A совпадают, хотя числа обусловленности собственных значений, вообще говоря, будут различны. Вычисленные последовательности $p_r(z)$ будут одинаковыми для всех таких $\tilde{A}$. Следовательно, если мы имеем, например, матрицу A вида

$$A = \begin{bmatrix} 0,2532 & 2^{40}(0,2615) & \\ 2^{-38}(0,3125) & 0,3642 & 0,1257 \\ & 0,2135 & 0,4132 \end{bmatrix}, \quad (9.1)$$

то при анализе ошибок мы можем использовать матрицу $\tilde{A}$ вида

$$\tilde{A} = \begin{bmatrix} 0,2532 & 2^1(0,2615) & \\ 2^1(0,3125) & 0,3642 & 0,1257 \\ & 0,2135 & 0,4132 \end{bmatrix}. \quad (9.2)$$

Конечно, мы должны ограничить исследуемые значения z областью $|z| \leq \|\tilde{A}\|_\infty$, а не $|z| \leq \|A\|_\infty$. В общем случае мы будем считать A уравновешенной матрицей (гл. 6, § 10).

10. Если трехдиагональная матрица была получена при помощи преобразования подобия из матрицы общего вида, то при использовании одинаковой точности вычислений в обоих случаях, ошибки в трехдиагональной матрице обычно будут больше, чем возмущения, эквивалентные ошибкам, сделанным при вычислении определителя.

В гл. 6, § 49 мы рекомендовали использовать арифметику с двойной точностью при приведении матрицы Хессенберга к трехдиагональному виду, даже если матрица Хессенберга была получена из матрицы общего вида с использованием обычной точности. Это было направлено против возможного возникновения «неустойчивых» стадий при приведении к трехдиагональному виду. Естественно поставить вопрос, можем ли мы вернуться к вычислениям с обычной точностью при вычислениях $p_r(z)$. Если при приведении к трехдиагональному виду не было неустойчивых стадий, то это сделать возможно, но если неустойчивые стадии были, то возвращение к работе с обычной точностью было бы ошибкой. Следующий элементарный анализ показывает, почему это так.

Мы видели в гл. 6, § 46, что в случае появления неустойчивости соответствующая трехдиагональная матрица имеет элементы, типичный порядок величин которых выражен в матрице

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 2^{-h} & 2^h \\ & & & & 2^h & 2^h & 2^{-h} \\ & & & & & 1 & 1 & 1 \\ & & & & & & 1 & 1 \end{bmatrix}, \quad (10.1)$$

но собственные значения которой порядка 1, а не 2^h , как можно предположить по величине нормы. Мы показали, что x и y , соответствующие правый и левый собственные векторы, нормированные так, что $y^T x = 1$, имеют компоненты порядка

$$\left. \begin{aligned} |y^T| &= (1, 1, 2^h, 2^h, 1, 1), \\ |x^T| &= (1, 1, 1, 1, 1, 1). \end{aligned} \right\} \quad (10.2)$$

Далее, наш анализ ошибок § 8 показал, что эквивалентное возмущение, соответствующее вычислению с использованием t -значной арифметики, имеет порядок величин

$$B = 2^{-t} \begin{bmatrix} 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 2^{-h} & 2^h & 2^h & \\ & & 2^h & 2^h & 2^{-h} \\ & & & 1 & 1 & 1 \\ & & & & 1 & 1 \end{bmatrix}. \quad (10.3)$$

Следовательно, из гл. 2, § 9, следует, что порядок величины возмущения собственного значения будет, вообще говоря, равен порядку величины $y^T B x / y^T x$, и мы имеем

$$y^T B x / y^T x = O(2^{2h-t}). \quad (10.4)$$

Это показывает, что при вычислении следует использовать ту же точность, что и при приведении. Однако, как показано в гл. 6, § 49, можно ожидать, что, вообще говоря, двойной точности будет достаточно, если только *исходная* матрица не слишком плохо обусловлена. Это значительно более удовлетворительная ситуация, чем в случае, связанном с формой Фробениуса.

Определители матриц Хессенберга

11. Наконец, мы рассмотрим вычисление $\det(A - zI)$ для фиксированного значения z , когда A — верхняя матрица Хессенберга. Снова находим, что это можно сделать, используя простые рекуррентные соотношения, которые выводятся следующим образом (Химан, 1957). Положим

$$P = A - zI \quad (11.1)$$

и обозначим столбцы P через $p_1, p_2, \dots, p_n$. Предположим, что мы можем найти скаляры $x_{n-1}, x_{n-2}, \dots, x_1$ такие, что

$$x_1 p_1 + x_2 p_2 + \dots + x_{n-1} p_{n-1} + p_n = k(z) e_1, \quad (11.2)$$

т. е. такие, что левая часть имеет нулевыми все компоненты, кроме первой. Тогда очевидно, что определитель матрицы Q , полученной из P заменой последнего столбца на $k e_1$, равен определителю P . Следовательно, раскладывая определитель Q по последнему столбцу, получим

$$\det(A - zI) = \det P = \det Q = \pm k(z) a_{21} a_{32} \dots a_{n, n-1}. \quad (11.3)$$

Элементы $a_{i+1, i}$ играют специальную роль. Мы подчеркнем это, используя обозначение

$$a_{i+1, i} = b_{i+1}. \quad (11.4)$$

Из уравнения (11.2) можно последовательно вычислить $x_{n-1}, x_{n-2}, \dots, x_1$, приравнявая элементы $n, n-1, \dots, 2$ обеих сторон. Тогда

$$b_r x_{r-1} + (a_{rr} - z) x_r + a_{r, r+1} x_{r+1} + \dots + a_{rn} x_n = 0 \quad (r = n, \dots, 2), \quad (11.5)$$

где $x_n = 1$. Окончательно, приравнявая первые компоненты, находим

$$(a_{11} - z) x_1 + a_{12} x_2 + \dots + a_{1n} x_n = k(z). \quad (11.6)$$

Определитель $(A - zI)$ отличается от $k(z)$ лишь на постоянный множитель, и, следовательно, можно ограничиться нахождением корней $k(z)$.

В каждом вычислении приблизительно $\frac{1}{2}n^2$ умножений. Если все поддиагональные элементы b_r равны единице, то x_{r-1} можно вычислять без использования деления. Мы видели в гл. 6, § 56, что если мы работаем с плавающей запятой, довольно удобно получать верхние матрицы Хессенберга, специализированные таким образом. Однако мы не будем делать никаких предположений о b_r , кроме того, что они не равны нулю. Если какой-либо из b_r равен нулю, можно работать с матрицами Хессенберга более низкого порядка.

Значения производных $f(z)$ можно также получать при помощи подобных рекуррентных соотношений. Дифференцируя (11.5) и (11.6) по z , получим

$$b_r x'_{r-1} + (a_{rr} - z) x'_r + a_{r, r+1} x'_{r+1} + \dots + a_{rn} x'_n - x_r = 0, \quad (11.7)$$

$$(a_{11} - z) x'_1 + a_{12} x'_2 + \dots + a_{1n} x'_n - x_1 = k'(z), \quad (11.8)$$

и так как $x_n = 1$, имеем $x'_n = 0$. Существуют подобные соотношения и для вторых производных. Производные x_i и $k(z)$ удобно поэтому вычислять одновременно с самими функциями.

Влияние ошибок округления

12. Вычисление x_i , определенных (11.5), почти необходимо проводить с плавающей запятой. Вычисляемая x_{r-1} выражается через вычисленные x_i ($i = r, \dots, n$) при помощи соотношения

$$\begin{aligned} -x_{r-1} &= fl[\{a_{rn}x_n + \dots + a_{r, r+1}x_{r+1} + (a_{rr} - z)x_r\}/b_r] \equiv \\ &\equiv \{a_{rn}x_n(1 + \varepsilon_{rn}) + \dots + a_{r, r+1}x_{r+1}(1 + \varepsilon_{r, r+1}) + \\ &\quad + (a_{rr} - z)x_r(1 + \varepsilon_{rr})\}(1 + \varepsilon_r)/b_r, \end{aligned} \quad (12.1)$$

где

$$\left. \begin{aligned} |\varepsilon_{ri}| &< (i - r + 2) 2^{-t_1} & (i > r), \\ |\varepsilon_{rr}| &< 3 \cdot 2^{-t_1}, & |\varepsilon_r| \leq 2^{-t}. \end{aligned} \right\} \quad (12.2)$$

Следовательно, мы имеем точные вычисления для матрицы с элементами a'_{ri} , b'_r равными

$$\left. \begin{aligned} a'_{ri} &= a_{ri}(1 + \varepsilon_{ri}), & |\varepsilon_{ri}| < (i - r + 2) 2^{-t_1} & (i > r), \\ a'_{rr} &= a_{rr}(1 + \varepsilon_{rr}) - z\varepsilon_{rr}, & |\varepsilon_{rr}| < 3 \cdot 2^{-t_1}, \\ b'_r &= b_r(1 + \eta_r), & |\eta_r| < 2^{-t_1}. \end{aligned} \right\} \quad (12.3)$$

Все эти элементы имеют малую относительную ошибку, кроме, быть может, a'_{rr} . Однако мы интересуемся только теми значениями z , для которых

$$|z| < \|A\|, \quad (12.4)$$

и, следовательно,

$$|a'_{rr} - a_{rr}| \leq \{|a_{rr}| + \|A\|\} |\varepsilon_{rr}|, \quad (12.5)$$

что показывает, что это возмущение мало по сравнению с любой нормой A .

Для иллюстрации рассмотрим случай, когда все элементы A по модулю ограничены единицей, так что

$$|z| \leq \|A\|_E \leq \left[\frac{1}{2} (n^2 + 3n - 2) \right]^{1/2}. \quad (12.6)$$

Тогда эквивалентное возмущение матрицы, являющееся функцией от z , равномерно ограничено матрицей F , которая при $n = 5$ имеет вид

$$F = 2^{-t_1} \begin{bmatrix} 3 & 3 & 4 & 5 & 6 \\ 1 & 3 & 3 & 4 & 5 \\ & 1 & 3 & 3 & 4 \\ & & 1 & 3 & 3 \\ & & & 1 & 3 \end{bmatrix} + 3\|A\|_E 2^{-t_1} I, \quad (12.7)$$

а так как $\|A\|_E = O(n)$, то

$$\|F\|_E \sim 2^{-t_1} (n^2/12) \quad \text{при } n \rightarrow \infty. \quad (12.8)$$

Снова мы находим, что ошибки, сделанные при вычислении, вряд ли будут серьезными, если только собственные значения не чрезвычайно чувствительны к вышеуказанным возмущениям элементов формы Хессенберга. Мы показали в главе 6, что существует несколько очень устойчивых методов приведения матрицы к форме Хессенберга. Следовательно, обычной точности, вообще говоря, достаточно и для приведения, и для последующих вычислений, если только исходная матрица не слишком плохо обусловлена.

Накопление с плавающей запятой

13. Хотя без сомнения результаты предыдущего параграфа вполне удовлетворительны, значительно лучшие оценки ошибок могут быть получены при вычислениях определителей матриц Хессенберга с использованием накопления с плавающей запятой. Для того чтобы получить все преимущества накопления, x_{r-1} следует вычислять из x_i ($i = r, \dots, n$) по соотношениям

$$-x_{r-1} = fl_2[(a_{rn}x_n + \dots + a_{r,r+1}x_{r+1} + a_{rr}x_r - zx_r)/b_r]. \quad (13.1)$$

Заметим, что мы написали $a_{rr}x_r - zx_r$, а не $(a_{rr} - z)x_r$, так как при вычислении $a_{rr} - z$ с одинарной точностью могут возникнуть ошибки округления. Уравнение (13.1) дает

$$-x_{r-1} = [a_{rn}x_n(1 + \varepsilon_{rn}) + \dots + a_{rr}x_r(1 + \varepsilon_r) - zx_r(1 + \eta_r)](1 + \varepsilon_r)/b_r, \quad (13.2)$$

где

$$|\varepsilon_{ri}| < 1, 5(i - r + 3)2^{-2t_2}, \quad |\eta_r| < 3 \cdot 2^{-2t_2}, \quad |\varepsilon_r| \leq 2^{-t}. \quad (13.3)$$

Поэтому вычисление является точным для матрицы с элементами a'_{ri} , b'_r , определяемыми соотношениями

$$\left. \begin{aligned} a'_{ri} &= a_{ri}(1 + \varepsilon_{ri}), & |\varepsilon_{ri}| &< 1, 5(i - r + 3)2^{-2t_2} \\ a'_{rr} &= a_{rr}(1 + \varepsilon_{rr}) - z\eta_r, & |\varepsilon_{rr}| &< (4, 5)2^{-2t_2}, & |\eta_r| &< 3 \cdot 2^{-2t_2}, \\ b'_r &= b_r(1 + \xi_r), & \xi_r &< 2^{-t_1}. \end{aligned} \right\} \quad (i = n, \dots, r + 1), \quad (13.4)$$

Снова для иллюстрации рассмотрим случай, когда все элементы A ограничены по модулю единицей. Эквивалентные возмущения теперь ограничены матрицей F , которая, например, при $n = 5$ имеет вид

$$F = (1,5) 2^{-2t_2} \begin{bmatrix} 3 & 4 & 5 & 6 & 7 \\ & 3 & 4 & 5 & 6 \\ & & 3 & 4 & 5 \\ & & & 3 & 4 \\ & & & & 3 \end{bmatrix} + 3 \|A\|_E 2^{-2t_2} I + 2^{-t_1} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \end{bmatrix}. \quad (13.5)$$

В общем случае три матрицы справа имеют 2-нормы, которые соответственно ограничены числами:

$$1,5(n+2)^2 2^{-2t_2}/12^{1/2}, \quad 3,0(n+2) 2^{-2t_2}/2^{1/2} \quad \text{и} \quad 2^{-t_1}.$$

Следовательно, предполагая, что $n^{2^{2-t}}$ значительно меньше единицы, получим, что $\|F\|_2$ порядка 2^{-t_1} , что является весьма удовлетворительным результатом.

Вычисление при помощи ортогональных преобразований

14. Метод вычисления Химана основан на умножении справа матрицы Хессенберга на элементарную матрицу типа S_n (гл. 1, § 40 (iv)), причем элементы S_n выбираются так, что последний столбец переходит в кратное e_1 . Было показано, что для численной устойчивости в прежних ситуациях существенно правило выбора ведущего элемента, но анализ ошибок §§ 12, 13 показал, что в нашем случае это не важно, и использование больших множителей x_i не дает вредного эффекта. Подобным же образом можно видеть, что, хотя мы и делим элементы на b_r , никакого вреда от этого не происходит, как бы малы они ни были. Можно ввести перестановки, но они в этом случае не принесут пользы. Способ, каким они могут быть введены, станет ясным в следующем параграфе, когда мы будем рассматривать использование плоских вращений.

Хотя боязнь численной неустойчивости здесь неосновательна, но все же естественна мысль об использовании плоских вращений (см., например, Уайт, 1958). Метод вычисления состоит из $(n-1)$ основного шага, причем на r -м шаге получается ноль в $(n-r+1)$ -й позиции последнего столбца. Конфигурация перед началом r -го шага в случае $n=5$, $r=3$ имеет вид

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times & 0 \\ & & & & \times & 0 \end{bmatrix}. \quad (14.1)$$

Основной r -й шаг состоит из умножения справа на вращение в плоскости $(n-r, n)$, причем угол вращения выбран так, что $a_{n-r+1, n}$ становится нулем. Шаг состоит в следующем:

- (i) Вычисляем $x = (b_{n-r+1}^2 + a_{n-r+1, n}^2)^{1/2}$.
- (ii) Вычисляем $\cos \theta = b_{n-r+1}/x$ и $\sin \theta = a_{n-r+1, n}/x$.
- (iii) Для всех j от 1 до $n-r$ вычисляем $a_{j, n-r} \cos \theta + a_{jn} \sin \theta$ и $-a_{j, n-r} \sin \theta + a_{jn} \cos \theta$ и записываем на места $a_{j, n-r}$ и a_{jn} соответственно.

(iv) Заменяем b_{n-r+1} на x и $a_{n-r, n}$ на нуль. Очевидно, что мы заменили столбцы $n-r$ и n их линейными комбинациями, так что нули, введенные на более ранних шагах, сохранились.

Окончательная матрица имеет ту же форму, что и матрица, соответствующая методу Химана, а определитель Δ равен

$$\Delta = \pm a_{1n} \prod_{r=1}^{n-1} b_{r+1}. \quad (14.2)$$

Обычным образом легко показать, что вычисленный определитель является точным определителем некоторой матрицы $(A + \delta A)$, где, например, для стандартных вычислений с плавающей запятой

$$\|\delta A\|_E \leq Kn^{3/2} \|A\|_E. \quad (14.3)$$

Поэтому метод устойчив, но оценка хуже, чем для метода Химана с стандартными вычислениями с плавающей запятой. Из соотношений (12.3)–(12.5) имеем

$$\|\delta A\|_E \leq Kn \|A\|_E, \quad (14.4)$$

однако для метода Химана, использующего накопление с плавающей запятой, мы имели лучший результат.

15. Так как в методе, использующем плоские вращения, в четыре раза больше умножений, можно безоговорочно утверждать, что во всех отношениях для этих вычислений лучше использовать элементарные преобразования без перестановок. В действительности метод Химана имеет некоторые другие преимущества, но они не видны сразу из нашего анализа. Мы показали, что вычисленный определитель является точным определителем для $(A + \delta A)$, и для каждого отличного от нуля элемента A получили оценку вида

$$|\delta a_{ij}| < 2^{-lf(i, j)} |a_{ij}|. \quad (15.1)$$

Рассмотрим теперь какое-либо преобразование подобия матрицы $(A + \delta A)$. Записав такое преобразование в виде $D^{-1}(A + \delta A)D = \tilde{A} + \delta \tilde{A}$, видим, что между элементами $\delta \tilde{A}$ и $\tilde{A}$ имеются те же соотношения, что и между элементами δA и A . Оценку влияния возмущения δA на собственные значения A можно рассматривать для самого хорошо обусловленного диагонального преобразования подобия матрицы A . Поэтому ясно, что перед применением метода Химана не нужно уравнивать A . Ни одно из этих замечаний не применимо к использованию ортогональных преобразований или элементарных преобразований подобия с перестановками.

Вычисление определителей матриц общего вида

16. Прежде чем рассматривать итерационные методы нахождения нулей, рассмотрим три вопроса, связанные с уже обсужденными. Первый касается вычисления $\det(A - zI)$, когда A — матрица общего вида. Как показано в главе 4, устойчивое определение получается при помощи любого приведения к треугольному виду, основанного на устойчивых элементарных преобразованиях или на ортогональных преобразованиях. Если мы можем накапливать скалярные произведения, то обычно предпочтительна прямая треугольная факторизация с перестановками.

Слабость таких методов заключается в большом объеме работы, а не в их устойчивости. Одно вычисление требует приблизительно $\frac{1}{3}n^3$ умножений, в то время как, если мы сначала приведем к форме Хессенберга, вычисление требует только $\frac{1}{2}n^2$ умножений.

Обобщенная проблема собственных значений

17. Второй вопрос касается решения системы n дифференциальных уравнений порядка r . В гл. 1, § 30 показано, что это приводит к обобщенной проблеме собственных значений, где требуется определить нули

$$\det[B_0 + B_1 z + \dots + B_{r-1} z^{r-1} - I z^r] = \det[C(z)]. \quad (17.1)$$

Хотя эта проблема может быть сведена к стандартной проблеме собственных значений, соответствующая матрица имеет порядок nr . Поэтому есть преимущества от работы прямо с формой (17.1) следующим образом.

Элементы $C(z)$ вычисляются для каждого значения z ; это требует приблизительно $n^2 r$ умножений. Определитель $C(z)$ затем может быть вычислен при помощи какого-либо устойчивого метода приведения к треугольному виду, описанного в главе 4. Наиболее экономный из этих методов требует $\frac{1}{3}n^3$ умножений. Так как каждый элемент $C(z)$ получен вычислением явного полинома, здесь есть потенциальная опасность, но на практике порядок исходных дифференциальных уравнений обычно не высок, чаще всего два или три. Обычно также B_i являются функциями параметра, и требуется исследовать зависимость отдельных собственных значений от параметра. Мы вернемся к этому вопросу в § 63.

Косвенное определение характеристического полинома

18. Третий вопрос состоит в том, что из-за простоты явной полиномиальной формы соблазнительно определять коэффициенты полинома по вычисленным значениям $\det(A - zI)$. Если $\det(A - zI)$, или некоторое постоянное кратное его, определены при $z_1, z_2, \dots, z_{n+1}$, то коэффициенты характеристического полинома являются решениями системы уравнений

$$c_n z_i^n + c_{n-1} z_i^{n-1} + \dots + c_1 z_i + c_0 = v_i \quad (i = 1, \dots, n+1), \quad (18.1)$$

где v_i это вычисленные значения. Аналогично, если требуются корни выражения (17.1), можно сделать $(nr + 1)$ вычислений определителя. Решению полиномиальных уравнений посвящено много усилий, и это является дальнейшим побуждением для приведения проблемы к этой форме.

Нашей главной целью здесь будет показать, что такие методы имеют присущие им жесткие ограничения. Для того чтобы проиллюстрировать это, рассмотрим один метод определения характеристического полинома и оценим его точность для одного специфического расположения нулей. Этот метод основан на том, что мы можем получить явное решение уравнений (18.1). Действительно,

$$c_{n-r} = \sum_{i=1}^{n+1} P_i^{(r)} v_i, \quad (18.2)$$

где

$$P_i^{(r)} = (-1)^r \sum_{i_k \neq i} z_{i_1} z_{i_2} \dots z_{i_r} / \prod_{j \neq i} (z_i - z_j). \quad (18.3)$$

Выбор z_i в нашем распоряжении. Если имеется некоторая оценка N нормы матрицы, то естественно взять $(n+1)$ равноотстоящих точек между $\pm N$. При таком выборе $P_i^{(r)}$ могут быть определены раз и навсегда для данного значения n с точностью до простого множителя N^r .

Предположим теперь, что собственные значения матрицы равны 1, 2, ..., 20, а наша оценка нормы оказалась равной 10. Тогда соответствующие значения z_i равны $-10, -9, \dots, +9, +10$. Для иллюстрации рассмотрим лишь определение коэффициента при z^{19} , причем сделаем предположение, что все величины v_i вычислены с максимальной относительной ошибкой 2^{-t} . В табл. 3 даны порядки величин множителей $P_i^{(1)}$ и величин v_i . Так как $v_i = 0$ ($i = 12, \dots, 21$), мы опустили соот-

4										Т а б л и ц а 3	
z_i	- 10	- 9	- 8	- 7	- 6	- 5	- 4	- 3	- 2	- 1	0
$ P_i^{(1)} $	2 ⁻⁵⁸	2 ⁻⁵⁴	2 ⁻⁵¹	2 ⁻⁴⁸	2 ⁻⁴⁶	2 ⁻⁴⁵	2 ⁻⁴⁴	2 ⁻⁴³	2 ⁻⁴³	2 ⁻⁴⁴	0
$ v_i $	286	285	283	281	279	277	275	272	269	266	262
$ P_i^{(1)} v_i $	2 ²⁸	2 ³¹	2 ³²	2 ³³	2 ³³	2 ³²	2 ³¹	2 ²⁹	2 ²⁶	2 ²²	0

ветствующие $P_i^{(1)}$, хотя по симметрии очевидно, что $P_i^{(1)} = -P_{n+2-i}^{(1)}$. Верное значение c_{19} равно 210 и, следовательно, порядка 2^8 , а мы видим, что отдельные слагаемые в сумме $\sum P_i^{(1)} v_i$ порядка 2^{33} . Поэтому при вычислении c_{19} должно иметь место значительное взаимное уничтожение. Ошибка в c_{19} , возникающая только от ошибки в $\det(A + 7I)$, будет порядка 2^{33-t} . Следовательно, даже при наших очень благоприятных предположениях, относительная ошибка в c_{19} вряд ли может быть меньше 2^{25-t} , а мы видели в § 5, что это фатально для точности вычисленных собственных значений, если только t не очень велико.

Полезным упражнением для читателя будет исследование ошибки в других коэффициентах и изучение эффекта изменения оценки нормы как для этого, так и для других расположений, обсуждавшихся в § 5. Вообще говоря, результаты, которые можно получить этим методом, совершенно неудовлетворительны. Исследовались различные модификации, включая, в частности, разложения по полиномам Чебышева, но результаты оказались неутешительными. Наш анализ показал, что даже если относительная ошибка в вычисленных величинах равна 2^{-t} , точные нули точного полинома, принимающего вычисленные значения, могут сильно отличаться от верных собственных значений.

Метод Лёверье

19. Независимый метод определения характеристического уравнения основан на замечании, что собственные значения A^r равны λ_i^r , и, следовательно, след A^r равен σ_r , где

$$\sigma_r = \sum_{i=1}^n \lambda_i^r. \quad (19.1)$$

Если мы вычислим σ_r ($r = 1, \dots, n$), то коэффициенты c_r могут быть получены из уравнений Ньютона

$$\left. \begin{aligned} c_{n-1} &= -\sigma_1, \\ r c_{n-r} &= -(\sigma_r + c_{n-1} \sigma_{n-1} + \dots + c_{n-r+1} \sigma_1), \quad (r = 2, \dots, n). \end{aligned} \right\} \quad (19.2)$$

Снова находим, что при вычислении c_r обычно происходит значительное взаимное уничтожение. Это может быть проверено, если рассмотреть порядки величин различных слагаемых в c_r . Рассмотрим, например, вычисление c_0 для матрицы, собственные значения которой равны 1, 2, ..., 20. Имеем:

$$c_0 = -(\sigma_{20} + c_{19}\sigma_{19} + \dots + c_1\sigma_1)/20. \quad (19.3)$$

Мы знаем, что c_0 равен $20!$, а первое слагаемое в правой части равно $-\sigma_{20}/20$, т. е. приблизительно $-20^{20}/21$. Следовательно, первое слагаемое противоположного c_0 знака и примерно в 2×10^6 раз больше. Некоторые следующие слагаемые даже еще больше, так что относительная ошибка 2^{-i} в некотором σ_i или c_i может вызвать значительно большую относительную ошибку в c_0 , вычисленном по ним.

Еще более поразительная иллюстрация получается, если рассмотреть расположение 2^{1-i} ($i = 1, \dots, 20$). Ясно, что все σ_i удовлетворяют соотношению

$$1 < \sigma_i < 2 \quad (19.4)$$

и что первое слагаемое в c_{n-r} равно $-\sigma_r/r$. Следующие c_i очень малы, так что взаимное уничтожение должно быть очень сильным. Мы не можем гарантировать даже то, что получим правильный знак этих коэффициентов, если только не используются очень точные вычисления.

Итерационные методы, основанные на интерполяции

20. Рассмотрим теперь методы, используемые для нахождения нулей. Сначала рассмотрим методы, основанные только на вычисленных значениях $f(z)$, а не на значениях ее производных. Эти методы являются в основном методами последовательной интерполяции. (Мы сейчас не делаем различия между интерполяцией и экстраполяцией.) В наиболее важной группе методов этого класса следующее приближение z_{k+1} к нулю определяется по $r+1$ предыдущим приближениям $z_k, z_{k-1}, \dots, z_{k-r}$ как нуль полинома степени r , проходящего через точки $(z_k, f(z_k)), \dots, (z_{k-r}, f(z_{k-r}))$. Из r таких нулей в качестве z_{k+1} берется тот, который лежит ближе всего к z_k . Если r равно 1 или 2, соответствующие полиномы линейные или квадратичные, и нахождение нулей тривиально. Удобно обозначить

$$f(z_i) = f_i. \quad (20.1)$$

В случае $r = 1$ имеем

$$z_{k+1} = z_k + \frac{f_k(z_k - z_{k-1})}{f_{k-1} - f_k} = \frac{z_k f_{k-1} - z_{k-1} f_k}{f_{k-1} - f_k}. \quad (20.2)$$

В случае $r = 2$ Мюллер (1956) предложил очень удобную схему вычислений. Положим

$$h_i = z_i - z_{i-1}, \quad \lambda_i = h_i/h_{i-1}, \quad \delta_i = 1 + \lambda_i. \quad (20.3)$$

Легко проверить, что требуемое значение λ_{k+1} удовлетворяет квадратному уравнению

$$\lambda_{k+1}^2 \lambda_k \delta_k^{-1} (f_{k-2} \lambda_k - f_{k-1} \delta_k + f_k) + \lambda_{k+1} \delta_k^{-1} g_k + f_k = 0, \quad (20.4)$$

где

$$g_k = f_{k-2} \lambda_k^2 - f_{k-1} \delta_k^2 + f_k (\lambda_k + \delta_k). \quad (20.5)$$

Отсюда получаем

$$\lambda_{k+1} = -2f_k \delta_k / \{g_k \pm [g_k^2 - 4f_k \delta_k \lambda_k (f_{k-2} \lambda_k - f_{k-1} \delta_k + f_k)]^{1/2}\}. \quad (20.6)$$

Знак в знаменателе выбирается так, чтобы соответствующее λ_{k+1} (и следовательно, h_{k+1}) было наименьшим по абсолютной величине.

Очевидно, что линейная интерполяция хороша своей простотой, но ее слабостью является то, что если z_k и z_{k-1} вещественны, и $f(z)$ — вещественная функция, то все последовательные значения вещественны. При квадратичной интерполяции мы можем выйти в комплексную плоскость, даже если мы начинаем с вещественных значений.

Если r больше двух, каждый шаг итерации требует решения полиномиального уравнения степени три или выше. Следовательно, вряд ли можно оправдать использование кубической или более высокой интерполяции, если только они не имеют значительно лучшие свойства сходимости (см. конец § 21).

Асимптотическая скорость сходимости

21. Рассмотрим предельную *) скорость сходимости в окрестности корня $z = \alpha$. Если положить

$$z - \alpha = w, \quad z_i - \alpha = w_i, \quad (21.1)$$

то

$$f_i = f_i(\alpha + w_i) = f'w_i + \frac{1}{2!}f''w_i^2 + \frac{1}{3!}f'''w_i^3 + \dots, \quad (21.2)$$

где через $f', f'', \dots$ обозначены производные при $z = \alpha$. Мы можем записать интерполяционный полином в терминах w . Полином порядка r , проходящий через точки (w_i, f_i) ($i = k, k-1, \dots, k-r$), имеет вид левой части уравнения

$$\det \begin{bmatrix} 0 & w^r & w^{r-1} & \dots & 1 \\ f_k & w_k^r & w_k^{r-1} & \dots & 1 \\ f_{k-1} & w_{k-1}^r & w_{k-1}^{r-1} & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ f_{k-r} & w_{k-r}^r & w_{k-r}^{r-1} & \dots & 1 \end{bmatrix} = 0. \quad (21.3)$$

Подставляя f_i из (21.2) и раскладывая по первой строке, получим

$$-\prod_{i < j} (w_i - w_j) \left[\frac{1}{r!} f^{(r)} w^r + \frac{1}{(r-1)!} f^{(r-1)} w^{r-1} + \dots + \frac{1}{1!} f' w + \right. \\ \left. + (-1)^r \frac{1}{(r+1)!} f^{(r+1)} w_k w_{k-1} \dots w_{k-r} \right] = 0, \quad (21.4)$$

где в коэффициент при каждом w^s включен только доминирующий член. Если w_i достаточно малы, требуемое решение w_{k+1} таково, что

$$w_{k+1} \sim (-1)^{r+1} f^{(r+1)} w_k w_{k-1} \dots w_{k-r} / [(r+1)! f']. \quad (21.5)$$

Вид этого решения оправдывает а posteriori то, что в (21.4) опущены все члены, кроме доминирующих.

*) Предельной скоростью сходимости автор называет скорость сходимости, начиная с достаточно хорошего приближения. (Прим. перев.)

Для оценки асимптотического поведения представим (21.5) в виде

$$|w_{k+1}| = K |w_k| |w_{k-1}| \dots |w_{k-r}|, \quad (21.6)$$

т. е.

$$(K^{1/r} |w_{k+1}|) = (K^{1/r} |w_k|) (K^{1/r} |w_{k-1}|) \dots (K^{1/r} |w_{k-r}|), \quad (21.7)$$

что дает

$$u_{k+1} = u_k + u_{k-1} + \dots + u_{k-r}, \quad u_i = \log(K^{1/r} |w_i|). \quad (21.8)$$

Из теории линейных разностных уравнений известно, что решением (21.8) будет

$$u_k = \sum c_i \mu_i^k, \quad (21.9)$$

где μ_i это корни уравнения

$$\mu^{r+1} = \mu^r + \mu^{r-1} + \dots + 1. \quad (21.10)$$

Это уравнение имеет один корень μ_1 между 1 и 2, причем

$$\mu_1 \rightarrow 2 \quad \text{при} \quad r \rightarrow \infty. \quad (21.11)$$

Другие r корней удовлетворяют уравнению $z^{r+2} - 2z^{r+1} + 1 = 0$, и все они лежат внутри единичного круга, что можно увидеть, применив теорему Руше к функциям

$$z^{r+2} + 1 \quad \text{и} \quad 2z^{r+1}. \quad (21.12)$$

Следовательно,

$$u_k \sim c_1 \mu_1^k, \quad K^{1/r} |w_k| \sim e^{c_1 \mu_1^k} \quad (21.13)$$

что дает $(K^{1/r} |w_{k+1}|) \sim (K^{1/r} |w_k|) \mu_1$,

$$|w_{k+1}| \sim K^{(\mu_1-1)/r} |w_k|^{\mu_1}. \quad (21.14)$$

При $r = 1$ имеем

$$\left. \begin{aligned} K &= |f''/2f|, \quad \mu_1 = \frac{1}{2} (1 + 5^{1/2}) \sim 1,62, \\ |w_{k+1}| &\sim K |w_k| |w_{k-1}|, \quad |w_{k+1}| \sim K^{0,62} |w_k|^{1,62}, \end{aligned} \right\} \quad (21.15)$$

а при $r = 2$

$$\left. \begin{aligned} K &= |-f'''/6f|, \quad \mu_1 \approx 1,84, \\ |w_{k+1}| &\sim K |w_k| |w_{k-1}| |w_{k-2}|, \quad |w_{k+1}| \sim K^{0,42} |w_k|^{1,84}. \end{aligned} \right\} \quad (21.16)$$

Асимптотическая скорость сходимости достаточно хороша в обоих случаях, но заметим, что большие значения K замедляют наступление предельной скорости сходимости. Так как μ_1 меньше двух при всех значениях r , то, по-видимому, мало доводов за использование интерполяционных полиномов высокой степени, по крайней мере при рассмотрении асимптотической скорости сходимости.

Кратные корни

22. В анализе предыдущего параграфа мы предполагали, что $f' \neq 0$, т. е. α — простой корень. Ниже приводится соответствующий анализ для случаев $r = 1$ и $r = 2$, когда α — двойной корень. В этом случае имеем

$$f_i = -\frac{1}{2!} f'' w_i^2 + \frac{1}{3!} f''' w_i^3 + \dots \quad (22.1)$$

При $r = 1$ линейное уравнение для w_{k+1} имеет вид

$$-\frac{1}{2} f'' (w_k^2 - w_{k-1}^2) w + \frac{1}{2} f'' w_k w_{k-1} (w_k - w_{k-1}) = 0, \quad (22.2)$$

что дает

$$w_{k+1} \sim w_k w_{k-1} / (w_k + w_{k-1}). \quad (22.3)$$

Если положить $u_i = 1/w_i$, то (22.3) станет

$$u_{k+1} = u_k + u_{k-1}. \quad (22.4)$$

Общее решение этого уравнения есть

$$u_k = a_1 \mu_1^k + a_2 \mu_2^k, \quad (22.5)$$

где μ_1 и μ_2 удовлетворяют уравнению

$$\mu^2 = \mu + 1. \quad (22.6)$$

Следовательно, окончательно имеем

$$w_k = 1/u_k \sim 1/(a_1 \mu_1^k) \approx 1/a_1 (1,62)^k, \quad (22.7)$$

так что в конечном счете ошибка делится приблизительно на 1,62 на каждом шаге.

При $r = 2$ квадратное уравнение для w_{k+1} имеет вид

$$\begin{aligned} & -(w_k - w_{k-1})(w_k - w_{k-2})(w_{k-1} - w_{k-2}) \times \\ & \times \left[\frac{1}{2} f'' w^2 - \frac{1}{6} f''' (w_k w_{k-1} + w_k w_{k-2} + w_{k-1} w_{k-2}) w + \right. \\ & \quad \left. + \frac{1}{6} f''' w_k w_{k-1} w_{k-2} \right] = 0, \end{aligned} \quad (22.8)$$

где снова в каждый коэффициент включены только доминирующие члены. В пределе решение дается соотношением

$$w_{k+1}^2 \sim (-f'''/3f'') w_k w_{k-1} w_{k-2}, \quad (22.9)$$

т. е.

$$|w_{k+1}|^2 = K |w_k| |w_{k-1}| |w_{k-2}|, \quad K = |-f'''/3f''|. \quad (22.10)$$

Мы установим это в конце параграфа. Положив $u_i = \log K |w_i|$, получим

$$2u_{k+1} = u_k + u_{k-1} + u_{k-2}. \quad (22.11)$$

Общее решение будет иметь вид

$$u_k = c_1 \mu_1^k + c_2 \mu_2^k + c_3 \mu_3^k, \quad (22.12)$$

где μ_1, μ_2, μ_3 — корни уравнения

$$2\mu^3 = \mu^2 + \mu + 1. \quad (22.13)$$

Как и в предыдущем параграфе, можно показать, что два корня этого уравнения лежат внутри единичного круга. Оставшийся μ_1 приблизительно равен 1,23. Следовательно, окончательно имеем

$$|w_{k+1}| \sim K^{\mu_1-1} |w_k|^{\mu_1} \approx |-f'''/3f''|^{0,23} |w_3|^{1,23}. \quad (22.14)$$

Этот результат оправдывает а posteriori то, что мы опустили из уравнения (22.8) для w ряд членов.

Обращение функционального соотношения

23. Если $f(z)$ имеет простой корень при $z = \alpha$, то z может быть выражено в виде сходящегося степенного ряда по f для достаточно малых f . Положим

$$z = g(f), \quad \text{где } z = \alpha, \quad \text{когда } f = 0. \quad (23.1)$$

Вместо приближений f полиномами по z мы можем приближать z полиномами по f (см., например, Островский (1960)). Общий случай достаточно хорошо иллюстрируется рассмотрением полиномов второй степени по f . Для приближающего полинома $L(f)$, проходящего через точки (z_k, f_k) , (z_{k-1}, f_{k-1}) , (z_{k-2}, f_{k-2}) , выраженного в форме Лагранжа, имеем

$$\begin{aligned} L(f) = z_k \frac{(f - f_{k-1})(f - f_{k-2})}{(f_k - f_{k-1})(f_k - f_{k-2})} + z_{k-1} \frac{(f - f_k)(f - f_{k-2})}{(f_{k-1} - f_k)(f_{k-1} - f_{k-2})} + \\ + z_{k-2} \frac{(f - f_k)(f - f_{k-1})}{(f_{k-2} - f_k)(f_{k-2} - f_{k-1})}. \end{aligned} \quad (23.2)$$

Классическая теория приближения полиномами Лагранжа дает для ошибки выражение

$$g(f) - L(f) = \frac{g'''(\eta)}{3!} (f - f_k)(f - f_{k-1})(f - f_{k-2}), \quad (23.3)$$

где η — некоторая точка интервала, содержащего f, f_k, f_{k-1}, f_{k-2} .

Взяв $f = 0$, получим

$$\alpha - L(0) = -\frac{g'''(\eta)}{3!} f_k f_{k-1} f_{k-2}. \quad (23.4)$$

Следовательно, если мы возьмем z_{k+1} равным $L(0)$, правая часть (23.4) даст нам выражение для ошибки. Из (23.2) имеем

$$\begin{aligned} z_{k+1} = L(0) = f_k f_{k-1} f_{k-2} \left[\frac{z_k}{f_k(f_k - f_{k-1})(f_k - f_{k-2})} + \right. \\ \left. + \frac{z_{k-1}}{f_{k-1}(f_{k-1} - f_k)(f_{k-1} - f_{k-2})} + \frac{z_{k-2}}{f_{k-2}(f_{k-2} - f_k)(f_{k-2} - f_{k-1})} \right]. \end{aligned} \quad (23.5)$$

Очевидно, что независимо от степени интерполяционного полинома, z_{k+1} единственно. Более того, если f — вещественная функция и мы

начали итерации с вещественных значений, то все z_i вещественны в противоположность нашей предыдущей технике, которая требует решения полиномиальных уравнений и, следовательно, делает возможным переход в комплексную плоскость.

Имеем

$$f_i \sim f(z_i) \approx (z_i - \alpha) f'(\alpha) \quad \text{при} \quad z_i \rightarrow \alpha. \quad (23.6)$$

Следовательно, (23.4) дает

$$(z_{k+1} - \alpha) \sim \frac{g'''(\eta)}{3!} [f'(\alpha)]^3 (z_k - \alpha) (z_{k-1} - \alpha) (z_{k-2} - \alpha), \quad (23.7)$$

и, положив $z_i - \alpha = w_i$, получим

$$w_{k+1} \sim K w_k w_{k-1} w_{k-2}, \quad \text{где} \quad K \approx g'''(\eta) [f'(\alpha)]^3 / 3!. \quad (23.8)$$

Это соотношение имеет ту же форму, что и (21.16), и показывает, что в конечном счете

$$|w_{k+1}| \sim K^{0,42} |w_k|^{1,84}. \quad (23.9)$$

В общем случае, когда используются интерполяционные полиномы степени r , получим

$$w_{k+1} = K w_k w_{k-1} \dots w_{k-r}, \quad \text{где} \quad K = -g^{(r+1)}(\eta) [f'(\alpha)]^{r+1} / (r+1)!. \quad (23.10)$$

Заметим, что если $r = 1$ этот метод совпадает с методом последовательной линейной интерполяции, обсуждавшимся в § 20. При r больше двух, получается мало преимуществ, так как мы видели в § 21, что в силу соотношения (23.10) предельная скорость сходимости никогда не достигает квадратичной, как бы велико ни было r .

Если α — корень кратности m , то

$$f = c_m (z - \alpha)^m + c_{m+1} (z - \alpha)^{m+1} + \dots, \quad (23.11)$$

и $z - \alpha$ можно представить в виде степенного ряда по степеням $f^{1/m}$. Очевидно, что в этом случае метод интерполяции будет значительно менее удовлетворителен.

По нашему опыту методы, описанные в этом параграфе, менее удовлетворительны для решения проблемы собственных значений, чем методы предыдущего параграфа. (С другой стороны, они оказались очень хороши для точного определения нулей функций Бесселя при условии хороших начальных приближений.) Поэтому в сравнении, которое мы будем делать в § 26, не будем обсуждать методов этого параграфа.

Метод деления отрезка пополам

24. Для того чтобы завершить описание методов, использующих только значения самой функции, мы должны упомянуть метод деления отрезка пополам. Этот метод уже обсуждался в главе 5 в связи с собственными значениями симметричных трехдиагональных матриц, но очевидно, что его можно использовать в более широком аспекте для нахождения вещественных нулей вещественных функций. Для его применения следует определить независимым образом такие значения a и b , что $f(a)$ и $f(b)$ будут разного знака. После того как это сделано, корень локализуется в интервале шириной $(b - a)/2^k$ за k шагов. Проблемы сходимости не возникает.

Метод Ньютона

25. Анализ §§ 21—23 касался только предельной скорости сходимости, и мы рассмотрели детально только случай корней кратности один и два. Кроме того, мы пренебрегали ошибками округления. Но, вообще говоря, трудно начать итерационный процесс с хороших приближений к корню, а вычисленные значения всегда будут подвержены влиянию ошибок округления. Эти вопросы будут рассмотрены в §§ 59, 60 и §§ 35—47 соответственно после изучения других итерационных методов.

Рассмотрим теперь методы, в которых в дополнение к значениям функций используются значения производных. Наиболее известным является метод Ньютона, в котором $(k+1)$ -е приближение связано с k -м соотношением

$$z_{k+1} = z_k - f(z_k)/f'(z_k). \quad (25.1)$$

В окрестности простого корня $z = \alpha$

$$w = z - \alpha, \quad f(z) = f(w + \alpha) = f'w + \frac{1}{2}f''w^2 + \dots, \quad (25.2)$$

где производные вычислены при $z = \alpha$. Следовательно, мы имеем

$$w_{k+1} = w_k - \frac{f'w_k + \frac{1}{2}f''w_k^2 + \frac{1}{6}f'''w_k^3 \dots}{f' + f''w_k + \frac{1}{2}f'''w_k^2 + \dots} = \quad (25.3)$$

$$= \frac{\frac{1}{2}f''w_k^2 + \frac{1}{3}f'''w_k^3 + \dots}{f' + f''w_k + \frac{1}{2}f'''w_k^2 + \dots}, \quad (25.4)$$

что дает

$$w_{k+1} \sim (f''/2f')w_k^2 \quad (f' \neq 0). \quad (25.5)$$

Для двойного корня $f' = 0$, $f'' \neq 0$, и, следовательно, из (25.3) следует

$$w_{k+1} = \frac{1}{2}w_k + \frac{\frac{1}{12}f'''w_k^3 + \dots}{f''w_k + \dots}, \quad (25.6)$$

что дает

$$w_{k+1} \sim \frac{1}{2}w_k. \quad (25.7)$$

Аналогично можно показать, что для корня кратности r

$$w_{k+1} \sim (1 - 1/r)w_k, \quad (25.8)$$

так что сходимость становится все медленнее вместе с ростом кратности. Очевидно, что для корня кратности r поправка, сделанная на каждом

шаге, уменьшается на множитель $1/r$. Поэтому для таких корней (25.1) заменяется на

$$z_{k+1} = z_k - rf(z_k)/f'(z_k), \quad (25.9)$$

и (25.9) дает

$$w_{k+1} \sim \frac{f^{(r+1)}}{r(r+1)f^{(r)}} w_k^2. \quad (25.10)$$

К сожалению, мы обычно не имеем *заранее* информации о кратности нулей.

Сравнение метода Ньютона с методом интерполяции

26. Вычисление первой производной для каждой из специальных форм матриц требует почти такого же объема работы, как и вычисление значений самой функции. Более того, если матрица достаточно высокого порядка, можем считать, что работа, связанная с одним шагом линейной или квадратичной интерполяции или метода Ньютона, мала по сравнению с вычислением значения функции или ее производной.

Поэтому естественно сравнивать два шага линейной или квадратичной интерполяции с одним шагом метода Ньютона. Для двух шагов линейной интерполяции в окрестности простого корня из (21.15) имеем

$$w_{k+2} \sim K^{0,62} (K^{0,62} w_k^{1,62})^{1,62} \approx K^{1,62} w_k^{2,62}, \quad \text{где } K = f''/2f, \quad (26.1)$$

а для метода Ньютона

$$w_{k+1} \approx K w_k^2 \quad (26.2)$$

с тем же значением K . Очевидно, что линейная интерполяция дает более высокую асимптотическую скорость сходимости.

В окрестности двойного корня для двух шагов линейной интерполяции

$$w_{k+2} \sim w_k/(1,62)^2 \approx w_k/2,62, \quad (26.3)$$

и снова получается выигрыш по сравнению с методом Ньютона. Методу Ньютона присущ один недостаток линейной интерполяции: если $f(z)$ — вещественная и мы начинаем с вещественного значения z_1 , то все z вещественные, и мы не можем определить комплексные нули.

Последовательная квадратичная интерполяция выглядит еще лучше при таком сравнении. Для простого корня два шага квадратичной интерполяции дают

$$|w_{k+2}| \sim K^{1,19} |w_k|^{3,39}, \quad \text{где } K = |-f'''/6f'|, \quad (26.4)$$

а для двойного корня

$$|w_{k+2}| \sim K^{0,51} |w_k|^{1,5}, \quad \text{где } K = |-f'''/3f''|. \quad (26.5)$$

Эти сравнения основаны на предположении, что вычисление значений производной сравнимо с вычислением значений функции. Для матриц общего вида, т. е. для функции $\det(A - zI)$, это наверняка неверно; вычисление значений производной требует значительно большей работы, чем вычисление значений самой функции. Контраст еще более

заметен для обобщенной проблемы собственных значений, где соответствующая функция имеет вид

$$\det (A_r z^r + A_{r-1} z^{r-1} + \dots + A_0).$$

Для такой проблемы чаша весов еще более склоняется в сторону интерполяционных методов.

Методы, дающие кубическую сходимость

27. Можно получить методы, имеющие асимптотическую сходимость любого порядка, если предположить, что мы готовы вычислять соответствующее число производных. Простой метод получения таких формул заключается в обращении разложения в ряд Тейлора. Имеем

$$f(z_k + h) = f(z_k) + hf'(z_k) + \frac{1}{2} h^2 f''(z_k) + O(h^3). \quad (27.1)$$

Следовательно, если h выбрано так, что $f(z_k + h) = 0$, то

$$\begin{aligned} z_{k+1} - z_k = h &= -f(z_k)/f'(z_k) - h^2 f''(z_k)/2f'(z_k) + O(h^3) \sim \\ &\sim -f(z_k)/f'(z_k) - [f(z_k)]^2 f''(z_k)/2[f'(z_k)]^3, \end{aligned} \quad (27.2)$$

и легко проверить, что это дает кубическую сходимость. Кажется, эта формула мало использовалась.

Метод Лагерра

28. Более интересная формула, также дающая кубическую сходимость в окрестности простого корня, принадлежит Лагерру, который развивал свой метод в связи с полиномами, имеющими только вещественные нули. Она основана на следующем.

Пусть $f(x)$ имеет вещественные нули $\lambda_1 < \lambda_2 < \dots < \lambda_n$, и пусть x лежит между λ_m и λ_{m+1} . Если x лежит вне интервала (λ_1, λ_n) , то вещественную ось можно рассматривать как замкнутую в бесконечно удаленной точке и говорить, что x лежит между λ_n и λ_1 . Рассмотрим теперь полином второй степени по X , определенный при всех вещественных u ,

$$g(X) = (x - X)^2 \left\{ \sum_{i=1}^n (u - \lambda_i)^2 / (x - \lambda_i)^2 \right\} - (u - X)^2. \quad (28.1)$$

Очевидно, что если $u \neq x$, то

$$g(\lambda_i) > 0 \quad (i = 1, \dots, n), \quad g(x) < 0. \quad (28.2)$$

Следовательно, для всех вещественных $u \neq x$ функция $g(X)$ имеет два нуля $X'(u)$ и $X''(u)$ таких, что

$$\lambda_m < X' < x < X'' < \lambda_{m+1}. \quad (28.3)$$

(Если x лежит вне интервала (λ_1, λ_n) , то X' лежит «между» λ_n и x , а X'' лежит «между» x и λ_1 .) Можно рассматривать X' и X'' как лучшие приближения двух корней, окружающих x , чем само x . Рассмотрим теперь наименьшее значение X' и наибольшее X'' , достижимые при всех возможных вещественных u . Вообще говоря, максимум и минимум будут достигаться при различных значениях u . Для экстремальных значений

имеем $\frac{\partial X'}{\partial u} = 0 = \frac{\partial X''}{\partial u}$. Если положить $g(X) = Au^2 + 2Bu + C$, где A, B, C — функции X , то для экстремальных значений корней получим

$$2Au + 2B = 0 \quad (28.4)$$

(так как $\frac{\partial A}{\partial u} = \frac{\partial B}{\partial u} = \frac{\partial C}{\partial u} = 0$, потому что $\frac{\partial X}{\partial u} = 0$), что дает $B^2 = AC$.

Легко проверить, что обращение в нуль дискриминанта действительно даст экстремальные значения.

Замечая, что

$$(f'/f) = \sum 1/(x - \lambda_i) = \sum_1, \quad (28.5)$$

$$(f'/f)^2 - (f''/f) = \sum 1/(x - \lambda_i)^2 = \sum_2, \quad (28.6)$$

так что $\sum \lambda_i/(x - \lambda_i)^2 = x \sum_2 - \sum_1$, мы видим после простых преобразований, что соотношение $B^2 = AC$ приводится к виду

$$n - 2 \sum_1 (x - X) + [\sum_1^2 - (n - 1) \sum_2] (x - X)^2 = 0. \quad (28.7)$$

Соответственно Лагерь дает два значения

$$\begin{aligned} X &= x - \frac{n}{\sum_1 \pm [n(n-1) \sum_2 - (n-1) \sum_1^2]^{1/2}} \\ &= x - \frac{nf}{f' \pm [(n-1)^2 (f')^2 - n(n-1) ff'']^{1/2}}, \end{aligned} \quad (28.8)$$

одно из которых ближе к λ_m , а другое к λ_{m+1} . Хотя обоснование применимо только к полиномам с вещественными нулями, можно сразу же распространить этот результат на комплексную плоскость, что приводит к следующему итерационному процессу:

$$z_{k+1} = z_k - nf_k/(f'_k \pm H_k^{1/2}), \quad \text{где } H_k = (n-1)^2 (f'_k)^2 - n(n-1) f_k f''_k. \quad (28.9)$$

В окрестности простого корня λ_m из (28.9) можно получить, что если мы выберем знак так, что знаменатель имеет наибольшую абсолютную величину, то

$$z_{k+1} - \lambda_m \sim \frac{1}{2} (z_k - \lambda_m)^3 [(n-1) \sum'_2 - (\sum'_1)^2]/(n-1), \quad (28.10)$$

где

$$\sum'_2 = \sum_{i \neq m} 1/(\lambda_m - \lambda_i)^2, \quad \sum'_1 = \sum_{i \neq m} 1/(\lambda_m - \lambda_i). \quad (28.11)$$

Поэтому в окрестности простого корня сходимость будет кубическая. Интересно, что при установлении кубического характера сходимости не используется то, что число n , входящее в формулу (28.9), есть степень f , так что процесс кубически сходится при всех n .

С другой стороны, если λ_m — корень кратности r , можно проверить, что (28.9) дает

$$z_{k+1} - \lambda_m \sim (r-1) (z_k - \lambda_m)/\{r + [(n-1)r/(n-r)]^{1/2}\}, \quad (28.12)$$

так что сходимость просто линейная.

Легко модифицировать (28.9) так, чтобы получить кубическую сходимость в окрестности корня кратности r . Соответствующая формула имеет вид

$$z_{k+1} = z_k - n f_k / \left[f'_k + \left\{ \frac{n-r}{r} [(n-1)(f'_k)^2 - n f_k f'_k] \right\}^{1/2} \right], \quad (28.13)$$

и она переходит в (28.9) при $r = 1$. К сожалению, нужно уметь находить кратность определяемого корня для того, чтобы использовать преимущества этой формулы.

Очевидно, что в случае, когда все корни вещественны, выбор определенного знака в формуле (28.9) дает монотонную последовательность, стремящуюся к одному из корней, в окрестности исходного приближения. Здесь нет аналогии с полиномами, имеющими комплексные корни.

Снова, если мы предположим, что вычисление значений f'' сравнимо с вычислением значений f и f' , два шага любого из этих кубически сходящихся процессов должны сравниваться с тремя шагами процесса Ньютона. Для двух шагов кубически сходящегося процесса имеем

$$w_{k+2} \sim A w_k^9, \quad (28.14)$$

а для трех шагов метода Ньютона

$$w_{k+3} \sim B w_k^8. \quad (28.15)$$

Поэтому метод Ньютона хуже, если рассматривать асимптотическое поведение.

Комплексные корни

29. До сих пор о комплексных корнях только упоминалось. Почти все методы, описанные выше как для вычисления значений $f(z)$ и значений ее производных, так и для вычисления корней, непосредственно переносятся на вычисление комплексных собственных значений комплексных матриц. Просто вещественная арифметика заменяется комплексной. Так как комплексное умножение требует четыре вещественных умножения, обычно в комплексном случае в четыре раза больше работы.

Однако комплексные матрицы на практике встречаются сравнительно редко. Наиболее часто встречается ситуация, когда матрица вещественна, но некоторые ее собственные значения суть комплексно сопряженные пары. Естественно, можно работать в этом случае, рассматривая все величины как комплексные, но можно попытаться для уменьшения объема вычислений получить некоторые преимущества от этого специального обстоятельства.

Заметим сначала, что если мы локализовали комплексный корень, то его комплексное сопряжение также будет корнем. Следовательно, предполагая, что объем вычислений при одной комплексной итерации превосходит не более чем вдвое объем при одной вещественной итерации, мы можем найти два комплексно сопряженных корня за то же время, что и два вещественных корня.

Если мы вычисляем явный полином для комплексного значения z по схеме Горнера (см. § 2), то все значения s_r , кроме s_n , комплексные, и на каждой стадии надо производить настоящие комплексные умножения. Поэтому для получения одного значения требуется в четыре раза

больше работы, и тот факт, что a_r вещественны, не дает преимуществ. В § 31 будет показано, как это обойти.

Для трехдиагональных матриц (ср. § 8), при¹

$$z = x + iy, \quad p_r(z) = s_r(z) + it_r(z), \quad (29.1)$$

уравнения (8.2) станут такими:

$$\left. \begin{aligned} s_r(z) &= (\alpha_r - x) s_{r-1}(z) + y t_{r-1}(z) - \beta_r \gamma_r s_{r-2}(z), \\ t_r(z) &= (\alpha_r - x) t_{r-1}(z) - y s_{r-1}(z) - \beta_r \gamma_r t_{r-2}(z). \end{aligned} \right\} \quad (29.2)$$

Таким образом, здесь требуется шесть умножений по сравнению с двумя в вещественном случае.

Наконец, для матриц Хессенберга (ср. § 11), при

$$x_r = u_r + iv_r, \quad z = x + iy, \quad (29.3)$$

уравнения (11.5) заменяется на такие:

$$\left. \begin{aligned} b_r u_{r-1} + (a_{rr} - x) u_r + a_{r, r+1} u_{r+1} + \dots + a_{rn} u_n + y v_r &= 0, \\ b_r v_{r-1} + (a_{rr} - x) v_r + a_{r, r+1} v_{r+1} + \dots + a_{rn} v_n - y u_r &= 0. \end{aligned} \right\} \quad (29.4)$$

С точностью до наличия двух дополнительных членов $y v_r$ и $-y u_r$ здесь как раз вдвое больший объем вычислений. Это же замечание приложимо к продифференцированным уравнениям (11.7). Следовательно, в этом случае мы можем находить комплексно сопряженные пары корней с приблизительно той же эффективностью, как и два вещественных корня, без существенной модификации процедуры.

30. Можно было бы подумать, что анализ ошибок, проведенный в §§ 12, 13, неприменим в комплексном случае, так как может случиться, что каждое из уравнений в (29.4) определяет независимо эквивалентное возмущение каждого a_{rs} и что, вообще говоря, они не будут одинаковыми. Однако, эта трудность иллюзорна, и ее можно преодолеть, допуская комплексные возмущения во всех a_{rs} . Для того чтобы избежать множества индексов, рассмотрим один отдельный элемент a_{rs} и предположим, что эквивалентные возмущения, соответствующие первому и второму уравнениям (29.4), равны соответственно $a_{rs} \varepsilon_1$ и $a_{rs} \varepsilon_2$. Мы хотим показать, что существуют такие η_1 и η_2 , что

$$a_{rs}(\eta_1 + i\eta_2)(u_s + iv_s) = a_{rs}u_s\varepsilon_1 + ia_{rs}v_s\varepsilon_2. \quad (30.1)$$

Очевидно,

$$(\eta_1^2 + \eta_2^2)(u_s^2 + v_s^2) = u_s^2\varepsilon_1^2 + v_s^2\varepsilon_2^2, \quad (30.2)$$

что дает

$$\eta_1^2 + \eta_2^2 \leq \varepsilon_1^2 + \varepsilon_2^2. \quad (30.3)$$

Можно показать, что оценки ошибок, данные для вычислений с плавающей запятой с вещественными величинами, переносятся на случай комплексных величин без каких-либо существенных модификаций. Соответствующие результаты таковы:

$$\left. \begin{aligned} fl(z_1 \pm z_2) &\equiv (z_1 + z_2)(1 + \varepsilon_1), & |\varepsilon_1| &\leq 2^{-t}, \\ fl(z_1 z_2) &\equiv z_1 z_2 (1 + \varepsilon_2), & |\varepsilon_2| &< 2 \cdot 2^{1/2} 2^{-t_1}, \\ fl(z_1/z_2) &\equiv z_1 (1 + \varepsilon_2)/z_2, & |\varepsilon_3| &< 5 \cdot 2^{1/2} 2^{-t_1}, \end{aligned} \right\} \quad (30.4)$$

где ε_i теперь, вообще говоря, комплексные. В этих результатах мы предполагаем, что ошибки округления в вещественной арифметике, которая необходима для этих комплексных вычислений, того же типа, который обсуждался в главе 3. Доказательство этих результатов оставлено для упражнений.

Комплексно сопряженные корни

31. Покажем теперь, что мы можем вычислить значение полинома с вещественными коэффициентами при комплексном значении аргумента приблизительно за $2n$ умножений. Заметим сначала, что $f(\alpha)$ есть остаток от деления $f(z)$ на $z - \alpha$. Отсюда следует, что $f(\alpha + i\beta)$ может быть получено как остаток при делении $f(z)$ на $z^2 - pz - l$, где

$$[z - (\alpha + i\beta)][z - (\alpha - i\beta)] = z^2 - pz - l. \quad (31.1)$$

В самом деле, если

$$f(z) \equiv (z^2 - pz - l)q(z) + rz + s, \quad (31.2)$$

то

$$f(\alpha + i\beta) = (r\alpha + s) + i(r\beta). \quad (31.3)$$

Если положить

$$\left. \begin{aligned} f(z) &= a_n z^n + a_{n-1} z^{n-1} + \dots + a_0, \\ q(z) &= q_n z^{n-2} + q_{n-1} z^{n-3} + \dots + q_2, \end{aligned} \right\} \quad (31.4)$$

то, приравнявая коэффициенты при одинаковых степенях z в обеих частях (31.2), получим после некоторых преобразований

$$\left. \begin{aligned} q_n &= a_n, & q_{n-1} &= a_{n-1} + pq_n, \\ q_s &= a_s + pq_{s+1} + lq_{s+2} & (s &= n-2, \dots, 2), \\ r &= a_1 + pq_2 + lq_3, & s &= a_0 + lq_2. \end{aligned} \right\} \quad (31.5)$$

Эти уравнения можно использовать для последовательного определения $q_n, q_{n-1}, \dots, q_2, r, s$. Заметим, что если мы продолжим рекуррентные соотношения так, что

$$q_1 = a_1 + pq_2 + lq_3, \quad (31.6)$$

$$q_0 = a_0 + pq_1 + lq_2, \quad (31.7)$$

то

$$r = q_1 \quad \text{и} \quad s = q_0 - pq_1 = q'_0. \quad (31.8)$$

Удобно переписать (31.2) в виде

$$f(z) \equiv (z^2 - pz - l)q(z) + q_1 z + q'_0. \quad (31.9)$$

При помощи рекуррентных соотношений и уравнений (31.3) можно вычислить $f(\alpha + i\beta)$ приблизительно за $2n$ вещественных умножений.

32. Подобный прием может быть применен для вычисления $f'(\alpha + i\beta)$. Для того чтобы сделать это, сначала разделим $q(z)$ на $z^2 - pz - l$. Если положить

$$q(z) \equiv (z^2 - pz - l)T(z) + T_1(z) + T'_0, \quad (32.1)$$

$$T(z) \equiv T_{n-2}z^{n-4} + T_{n-3}z^{n-5} + \dots + T_2, \quad (32.2)$$

то аналогичными рассуждениями получим

$$\left. \begin{aligned} T_{n-2} &= q_n, & T_{n-3} &= q_{n-1} + pT_{n-2}, \\ T_s &= q_{s+2} + pT_{s+1} + lT_{s+2} & (s = n-4, \dots, 0), \\ T'_0 &= T_0 - pT_1. \end{aligned} \right\} \quad (32.3)$$

Если эти две системы рекуррентных соотношений используются, то нам не надо запоминать промежуточные q_s . Дифференцируя (31.9), получим

$$f'(z) = (z^2 - pz + l)q'(z) + (2z - p)q(z) + q_1, \quad (32.4)$$

$$f'(\alpha + i\beta) = (2\alpha + 2i\beta - p)q(\alpha + i\beta) + q_1, \quad (32.5)$$

и из (32.1) найдем

$$q(\alpha + i\beta) = T_1\alpha + T'_0 + iT_1\beta. \quad (32.6)$$

Следовательно, вспоминая, что $2\alpha = p$, получим

$$f'(\alpha + i\beta) = 2i\beta(T_1\alpha + T'_0 + iT_1\beta) + q_1. \quad (32.7)$$

Аналогично мы можем получить $f''(\alpha + i\beta)$, если мы поделим $T(z)$ на $z^2 - pz - l$, и т. д.

Метод Берстоу

33. В предыдущем параграфе мы ввели деление на $z^2 - pz - l$, чтобы вычислить значение $f(z)$ для комплексного значения аргумента. Заметим однако, что рекуррентные соотношения для q_r позволяют нам вычислить $f(z)$ при z_1 и z_2 , корнях полинома $z^2 - pz - l$, независимо от того, являются ли они комплексно сопряженной парой или вещественными.

В методе Берстоу (1914) производится итерация прямо для вещественного квадратичного множителя $z^2 - pz - l$. Дифференцируя (31.9) по l и p соответственно, получаем

$$0 = -q(z) + (z^2 - pz - l) \frac{\partial}{\partial l} q(z) + \frac{\partial q_1}{\partial l} z + \frac{\partial q'_0}{\partial l}, \quad (33.1)$$

$$0 = -zq(z) + (z^2 - pz - l) \frac{\partial}{\partial p} q(z) + \frac{\partial q_1}{\partial p} z + \frac{\partial q'_0}{\partial p}, \quad (33.2)$$

что можно записать в виде

$$\frac{\partial q_1}{\partial l} z + \frac{\partial q'_0}{\partial l} \equiv q(z) \pmod{z^2 - pz - l}, \quad (33.3)$$

$$\frac{\partial q_1}{\partial p} z + \frac{\partial q'_0}{\partial p} \equiv zq(z) \pmod{z^2 - pz - l}. \quad (33.4)$$

Следовательно, из (32.1) получаем

$$\frac{\partial q_1}{\partial l} z + \frac{\partial q'_0}{\partial l} \equiv T_1 z + T'_0, \quad (33.5)$$

$$\frac{\partial q_1}{\partial p} z + \frac{\partial q'_0}{\partial p} \equiv z(T_1 z + T'_0) \equiv z(pT_1 + T'_0) + lT_1. \quad (33.6)$$

Используя метод Ньютона для исправления приближенного нуля $q_1 z + q'_0$, выбираем δp и δl так, что

$$(q_1 z + q'_0) + \frac{\partial}{\partial p} (q_1 z + q'_0) \delta p + \frac{\partial}{\partial l} (q_1 z + q'_0) \delta l = 0 \quad (33.7)$$

при всех z и, следовательно,

$$q_1 + (p T_1 + T_0) \delta p + T_1 \delta l = 0, \quad (33.8)$$

$$q'_0 + l T_1 \delta p + T_0 \delta l = 0. \quad (33.9)$$

Решение может быть записано в виде

$$\left. \begin{aligned} D \delta p &= T_1 q_0 - T_0 q_1, & D \delta l &= M q_1 - T_0 q_0, \\ M &= l T_1 + p T_0, & D &= T_0^2 - M T_1. \end{aligned} \right\} \quad (33.10)$$

где

Так как в (33.7) были опущены члены δp^2 , $\delta p \delta l$, δl^2 и выше, этот метод дает квадратичную сходимость к квадратичному множителю. Даже в случае сходимости к квадратичному множителю с комплексными корнями, метод Берстоу не совпадает с методом Ньютона. В частности, если он применяется к полиномам второй степени, то он дает сходимость к верному множителю с произвольного начального приближения в одну итерацию. Метод Ньютона дает просто квадратичную сходимость.

Обобщенный метод Берстоу

34. Обычно метод Берстоу обсуждают в связи с явными полиномами. Однако мы можем применить его к полиномам, выраженным в любой форме, предполагая, что можем вычислить остатки, соответствующие двум последовательным делениям на $z^2 - pz - l$.

Для трехдиагональной матрицы можно получить простое рекуррентное соотношение для остатков, если главные ведущие миноры $(A - zI)$ делятся на $z^2 - pz - l$. Если мы напомним

$$\left. \begin{aligned} p_r(z) &\equiv (z^2 - pz - l) q_r(z) + A_r z + B_r, \\ q_r(z) &\equiv (z^2 - pz - l) T_r(z) + C_r z + D_r, \end{aligned} \right\} \quad (34.1)$$

то из соотношения

$$p_r(z) = (\alpha_r - z) p_{r-1}(z) - \beta_r \gamma_r p_{r-2}(z) \quad (34.2)$$

имеем, подставляя $p_i(z)$ из (34.1) и приравнявая коэффициенты при $z^2 - pz - l$,

$$q_r(z) = (\alpha_r - z) q_{r-1}(z) - \beta_r \gamma_r q_{r-2}(z) - A_{r-1}, \quad (34.3)$$

и аналогично

$$T_r(z) = (\alpha_r - z) T_{r-1}(z) - \beta_r \gamma_r T_{r-2}(z) - C_{r-1}. \quad (34.4)$$

Эти уравнения дают

$$\left. \begin{aligned} A_r &= (\alpha_r - p) A_{r-1} - \beta_r \gamma_r A_{r-2} - B_{r-1}, \\ B_r &= \alpha_r B_{r-1} - \beta_r \gamma_r B_{r-2} - l A_{r-1}, \\ C_r &= (\alpha_r - p) C_{r-1} - \beta_r \gamma_r C_{r-2} - D_{r-1}, \\ D_r &= \alpha_r D_{r-1} - \beta_r \gamma_r D_{r-2} - l C_{r-1} - A_{r-1}. \end{aligned} \right\} \quad (34.5)$$

Так как $p_0(z) = 1$, $q_0(z) = 0$, $p_1(z) = \alpha_1 - z$, $q_1(z) = 0$, то

$$A_0 = 0, \quad B_0 = 1, \quad A_1 = -1, \quad B_1 = \alpha_1, \quad C_0 = D_0 = C_1 = D_1 = 0. \quad (34.6)$$

Заметим, что первые два уравнения (34.5) позволяют нам определить $p_n(\alpha + i\beta)$ за $5n$ умножений, сравнительно с $6n$ умножениями, необходимыми при использовании уравнений (29.2). Величины A_n , B_n , C_n , D_n будут соответственно q_1 , q'_0 , T_1 , T'_0 из предыдущего параграфа.

Мы можем получить соответствующие соотношения для ведущих главных миноров $(A - zI)$, если A — верхняя матрица Хессенберга. Для упрощения обозначений предположим, что все поддиагональные элементы равны единице. Тогда имеем

$$p_r(z) = (a_{rr} - z)p_{r-1}(z) - a_{r-1,r}p_{r-2}(z) + \\ + a_{r-2,r}p_{r-3}(z) + \dots + (-1)^{r-1}a_{1r}p_0(z), \quad (34.7)$$

где $p_0(z) = 1$.

Определяя $q_r(z)$, A_r , B_r , C_r , D_r , как в (34.1), мы видим, что $q_r(z)$ удовлетворяют соотношению

$$q_r(z) = (a_{rr} - z)q_{r-1}(z) - a_{r-1,r}q_{r-2}(z) + \\ + a_{r-2,r}q_{r-3}(z) + \dots + (-1)^{r-1}a_{1r}q_0(z) - A_{r-1}, \quad (34.8)$$

а A_r , B_r , C_r , D_r — соотношениям

$$\left. \begin{aligned} A_r &= (a_{rr} - p)A_{r-1} - a_{r-1,r}A_{r-2} + a_{r-2,r}A_{r-3} + \dots \\ &\quad \dots + (-1)^{r-1}a_{1r}A_0 - B_{r-1}, \\ B_r &= a_{rr}B_{r-1} - a_{r-1,r}B_{r-2} + a_{r-2,r}B_{r-3} + \dots \\ &\quad \dots + (-1)^{r-1}a_{1r}B_0 - lA_{r-1}, \\ C_r &= (a_{rr} - p)C_{r-1} - a_{r-1,r}C_{r-2} + a_{r-2,r}C_{r-3} + \dots \\ &\quad \dots + (-1)^{r-1}a_{1r}C_0 - D_{r-1}, \\ D_r &= a_{rr}D_{r-1} - a_{r-1,r}D_{r-2} + a_{r-2,r}D_{r-3} + \dots \\ &\quad \dots + (-1)^{r-1}a_{1r}D_0 - lC_{r-1} - A_{r-1}, \end{aligned} \right\} \quad (34.9)$$

где

$$A_0 = 0, \quad B_0 = 1, \quad A_1 = -1, \quad B_1 = \alpha_1; \quad C_0 = D_0 = C_1 = D_1 = 0. \quad (34.10)$$

Для каждой итерации требуется приблизительно $2n^2$ умножений. Хандскомб (1962) дал несколько отличную формулировку рекуррентных соотношений.

Практические рассмотрения

35. До настоящего времени при обсуждении итерационных методов мы не дали никаких указаний на трудности вычислений. Наши обсуждения асимптотического поведения сходимости были бы реалистичными, если бы собственные значения вычислялись с громадным количеством значащих цифр. На практике редко интересуются более чем 10 десятичными знаками, хотя чтобы получить их, нужно работать со значительно большей точностью.

Обычным способом описания поведения методов, имеющих квадратичную или кубическую сходимость, является утверждение, что «за каждую итерацию число значащих цифр удваивается или утраивается». Однако такие утверждения мог бы делать «классический» аналитик, но они неприменимы в практическом численном анализе. Предпопо-

жим, например, что при использовании кубически сходящегося процесса предельное поведение погрешности w_i описывается соотношением

$$w_{k+1} \approx 10^4 w_k^3. \quad (35.1)$$

Если $w_k = 10^{-3}$, то

$$w_{k+1} \approx 10^{-5}, \quad w_{k+2} \approx 10^{-11}, \quad w_{k+3} \approx 10^{-29}, \quad (35.2)$$

предполагая, что асимптотическое соотношение уже верно при $w_k = 10^{-3}$. Но на практике мы не можем ожидать, что $w_{k+3} = 10^{-29}$, если только мы не используем более 29 десятичных знаков для наших значений функции. Итерации, скорее всего, будут остановлены на w_{k+2} , и если мы будем получать w_{k+3} , то точность будет, вероятно, ограничена точностью вычислений. Это верно, даже если вычисленные значения $f(z)$ и ее производных верны с рабочей точностью. Почти сразу после того, как мы достигаем значения w_k , которое достаточно мало для того, чтобы асимптотическое поведение вступило в силу, мы находим, что точность вычислений ограничивает наши возможности получения этого преимущества.

Достижимая точность далее лимитируется нашей способностью вычисления значений самой $f(z)$ в окрестности корня. Исследование формул, которые мы дали, показывает, что в каждом случае «поправка» $z_{k+1} - z_k$ прямо пропорциональна текущему вычисленному значению $f(z_k)$. Вычисленная поправка поэтому не будет правильной поправкой, если только $f(z_k)$ не имеет несколько правильных значащих цифр.

Влияние ошибок округления на асимптотическую сходимость

36. Мы нашли для каждой из специальных форм матриц оценки для возмущений δA матрицы A , которые эквивалентны ошибкам, сделанным при вычислении $\det(A - zI)$. В каждом случае оценка для $\|\delta A\|_E$ пропорциональна 2^{-t} , и, следовательно, возмущение может быть сделано столь угодно малым, если работать с достаточно высокой точностью. Нужно подчеркнуть, что δA является функцией z и алгоритма, используемого при вычислении $\det(A - zI)$.

Если обозначить собственные значения A через λ_i , а $(A + \delta A)$ через λ'_i , то из непрерывности собственных значений следует, что их можно упорядочить так, что

$$\lambda'_i \rightarrow \lambda_i \quad \text{при} \quad t \rightarrow \infty.$$

Однако пока t не достаточно велико, собственные значения $(A + \delta A)$ могут иметь мало общего с собственными значениями A . Хороший пример этого дает форма Фробениуса с собственными значениями 1, 2, ..., 20. Мы видели (уравнение (5.3)), что пока 2^{-t} не станет меньше, чем 10^{-14} , собственные значения возмущенной матрицы вряд ли можно рассматривать как собственные значения исходной матрицы.

Метод деления отрезка пополам

37. Рассмотрим сначала ограничения для метода деления отрезка пополам, который используется для локализации простого вещественного собственного значения λ_p вещественной матрицы. Предположим, что мы нашли две точки a и b , в которых $\det(A - zI)$ имеет противоположные знаки. Соответственно нашему вычислению $\det(A - zI)$ при некотором

значении z мы нашли $\det(A + \delta A - zI)$, где δA — вещественная, и

$$\det(A + \delta A - zI) = \prod_{i=1}^n (\lambda'_i - z). \quad (37.1)$$

Для каждой из специальных форм мы дали оценки для δA , и для всех допустимых δA каждое соответствующее λ'_i будет лежать в связной области, содержащей λ_i . Можно назвать эту область *областью неопределенности*, связанной с λ_i . Заметим, что эта область зависит от t и алгоритма, используемого для вычисления $\det(A - zI)$. Предположим, что t достаточно велико для того, чтобы область, содержащая λ_p , была изолирована от остальных областей. В этом случае, конечно, все λ'_p должны быть вещественными, а область, содержащая λ_p , это интервал вещественной оси.

Типичная ситуация показана на рис. 4. Определяемое собственное значение есть λ_5 , оно остается вещественным при всех возмущениях δA ,

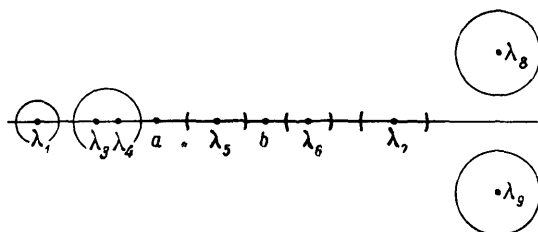


Рис. 4.

удовлетворяющих соответствующему условию. Остальные собственные значения ведут себя следующим образом.

λ_1 — двойное собственное значение, которое становится комплексно сопряженной парой при некоторых δA .

λ_3 и λ_4 — близкие собственные значения, которые становятся комплексно сопряженной парой при некоторых δA .

λ_6 и λ_7 — вещественные собственные значения, которые остаются вещественными.

λ_8 и λ_9 — комплексно сопряженная пара, которая остается комплексно сопряженной парой.

Диаметр области, содержащей λ_i , является мерой его обусловленности.

Пусть $\det(A - zI)$ вычислялся в точках a и b , и мы получили положительное и отрицательное значения соответственно, независимо от расположения λ'_i в их областях. Если мы теперь будем делить отрезок (a, b) , то будем получать верные знаки $\det(A - zI)$ для любых точек деления, не лежащих в области, содержащей λ_5 . Следовательно, точно так же как в гл. 5, § 41, можно показать, что либо все интервалы деления действительно содержат λ_5 , либо, начиная с некоторой стадии, по крайней мере одна из граничных точек интервалов лежит в области, содержащей λ_5 . Если эта область есть интервал $(\lambda_5 - x, \lambda_5 + y)$, то очевидно, что после k шагов деления центр окончательного интервала деления лежит в пределах

$$[\lambda_p - x - 2^{-k-1}(b - a), \lambda_p + y + 2^{-k-1}(b - a)]. \quad (37.2)$$

Так как мы можем использовать произвольное число шагов, окончательным ограничением метода того является размер интервала $(\lambda_5 - x, \lambda_5 + y)$.

Заметим, что мы вообще не можем ожидать, что вычисленные значения малых собственных значений будут иметь малую относительную ошибку. Нет никаких оснований ожидать, что области, содержащие малые корни, будут меньше, чем области, содержащие большие корни. Вообще, в лучшем случае, мы можем ожидать, что область будет порядка $2^{-t} \|A\|_E$, если собственное значение хорошо обусловлено.

Последовательная линейная интерполяция

38. Проблема сходимости не возникает в методе деления пополам, и можно понять, что успех этого метода непосредственно зависит от того, что мы никак не использовали величину значения функции. Например, если точка деления z лежит вне области, содержащей некоторое собственное значение λ_p , то знак вычисленного $\det(A - zI)$ будет, конечно, правильным, но относительная ошибка может быть сколь угодно велика, как можно видеть из (37.4) и того, что λ'_p может быть всюду в λ_p -области.

С другой стороны, если использовать последовательную линейную интерполяцию, то

$$z_{k+1} = z_k + f_k(z_k - z_{k-1})/(f_{k-1} - f_k), \quad (38.1)$$

и, на первый взгляд, высокая относительная ошибка в f_k может оказаться серьезной. Однако так только кажется. Если точность вычислений такова, что можно серьезно рассматривать проблему локализации λ_p с разумной точностью, то на стадии, когда мы достигнем z_k , которое близко к области λ_p , z_k и z_{k-1} будут лежать по одну сторону от λ_p , и, как следует из § 21, $|z_k - \lambda_p|$ будет значительно меньше, чем $|z_{k-1} - \lambda_p|$. Если λ'_i и λ''_i суть собственные значения возмущенной матрицы $(A + \delta A)$, отвечающей вычислениям z_k и z_{k-1} , то имеем

$$\begin{aligned} z_{k+1} &= z_k + (z_k - z_{k-1}) \prod_{i=1}^n (\lambda'_i - z_k) / \left[\prod_{i=1}^n (\lambda''_i - z_{k-1}) - \prod_{i=1}^n (\lambda'_i - z_k) \right] \approx \\ &\approx z_k + (z_k - z_{k-1}) (\lambda'_p - z_k) / [(\lambda''_p - z_{k-1}) - (\lambda'_p - z_k)], \end{aligned} \quad (38.2)$$

так как

$$\lambda'_i - z_k \approx \lambda''_i - z_k \quad (i \neq p). \quad (38.3)$$

Соотношение (38.2) может быть записано в виде

$$z_{k+1} \approx z_k + (\lambda'_p - z_k) \left/ \left[1 + \frac{\lambda''_p - \lambda'_p}{z_k - z_{k-1}} \right] \right., \quad (38.4)$$

откуда ясно, что z_{k+1} очень близко к λ'_p . Вообще говоря, z_{k+1} будет значительно менее точным, чем оно было бы получено при точных вычислениях; последнее таково, что $z_{k+1} - \lambda_p$ порядка $(z_k - \lambda_p)(z_{k-1} - \lambda_p)$, как это показано в (21.15). Тем не менее можно ожидать, что итерации будут улучшаться до тех пор, пока мы не получим значение, лежащее в области неопределенности, и даже используя метод деления пополам, мы не могли бы достичь лучшего.

После того как мы достигли области неопределенности, ситуация ухудшается. Вычисленные значения f почти точно пропорциональны $z_k - \lambda_p$, а λ'_p меняются с изменением z_k . Следовательно, для z_k внутри области неопределенности вычисленные значения $\det(A - z_k I)$ будут колебаться по причудам ошибок округления и «ожидаемые» значения будут более или менее постоянными для всех таких z_k . Это хорошо видно

в табл. 2 § 7, где первые пять значений z_k лежат в области неопределенности, окружающей $\lambda_p = 10$. Несмотря на то, что расстояние пятого значения от корня в 2^{19} раз больше подобного расстояния для первого, все вычисленные значения имеют один порядок величины. Из уравнения (38.1) видно, что поправка при одной итерации равна

$$(z_k - z_{k-1})[f_k/(f_{k-1} - f_k)].$$

Для значений z_k и z_{k-1} внутри области неопределенности величина в квадратных скобках целиком зависит от ошибок округления. Вообще говоря, эта величина будет порядка единицы (это верно, например, для всех пяти значений в табл. 2), но может случиться, что f_{k-1} и f_k исключительно близки, так что эта величина будет очень велика.

39. Соответственно поведение итераций можно описать следующим образом. На значительном расстоянии от собственного значения вычисленные значения функции имеют малую относительную ошибку, и поправка, получаемая во время типичного шага, в основном такая же, как при точных вычислениях. При приближении к области неопределенности относительная ошибка вычисленных значений возрастает, хотя пока текущие значения вне области неопределенности, движение происходит в правильном направлении. Заметный прогресс прекращается, когда мы попадаем внутрь области неопределенности. После этого последовательные приближения стремятся оставаться внутри или около области неопределенности, хотя при неблагоприятной комбинации ошибок округления в двух последовательных вычислениях может получиться значение, существенно отстоящее от этой области.

Полезно провести соответствующие вычисления для пар значений z_k из табл. 2, используя как вычисленные, так и «верные» значения. Напомним, что даже эти так называемые верные значения не являются точными, так как при их представлении малым количеством значащих цифр делалась ошибка округления. Важно понять, что если, скажем, нужно определить собственное значение при $z = 10$ с точностью до s двоичных знаков, то если $|z_k - 10|$ порядка 2^{-p} , достаточно вычислять $f(z_k)$ с $s-p$ верными двоичными знаками. При этом поправка, делаемая на каждом шаге, совпадает с поправкой, соответствующей точным вычислениям, во всех рассматриваемых $s-p$ цифрах.

Так, в табл. 2 значение при $z = 10 + 2^{-28} + 7 \times 2^{-42}$ дано приблизительно с 15 значащими двоичными цифрами. Если мы возьмем это значение и выполним один шаг линейной интерполяции, используя $z = 10 + 2^{-18} + 7 \times 2^{-42}$ в качестве второй точки интерполяции, интерполированное значение будет иметь 43 правильных двоичных цифры, т. е. $28 + 15$.

Кратные и патологически близкие собственные значения

40. Метод деления неприменим для локализации кратных нулей четного порядка, но использовать последовательную линейную интерполяцию все еще можно. Можно подумать, что точность, достижимая в этом случае для кратного корня, сильно ограничена, но это не так. Хотя сходимость медленная, окончательная достижимая точность зависит только от размеров области неопределенности. Ситуация теперь значительно хуже, чем в случае простого нуля. Для двойного нуля, например, соотношение (38.2) становится таким:

$$z_{k+1} \approx z_k + (z_k - z_{k-1}) (\lambda'_p - z_k)^2 / [(\lambda''_p - z_{k-1})^2 - (\lambda'_p - z_k)^2], \quad (40.1)$$

но мы не можем утверждать, что $z_k - \lambda_p$ значительно меньше, чем

$z_k \rightarrow \lambda_p$, так как даже при точных вычислениях погрешность уменьшается только в 1,62 раза за итерацию (ср. (22.7)). Поэтому возможна неустойчивость, когда $z_k - \lambda_p$ в несколько раз превышает диаметр области неопределенности, если случайно λ'_p и λ''_p лежат по разные стороны λ_p .

Если кратное собственное значение плохо обусловлено или даже если оно хорошо обусловлено, но метод вычисления плохо обусловлен, то область неопределенности будет сравнительно велика. Например, если мы используем явную форму функции $z^3 - 3z^2 + 3z - 1$ и значение $f(z)$ вычисляем по схеме Горнера, то мы знаем, что каждое вычисление дает точное значение некоторого полинома

$$(1 + \varepsilon_3)z^3 - 3(1 + \varepsilon_2)z^2 + 3(1 + \varepsilon_1)z - (1 + \varepsilon_4), \quad (40.2)$$

где границы для ε_i порядка 2^{-i} . Такие возмущения коэффициентов иногда вызывают в корнях возмущения порядка $2^{-i/3}$ (гл. 2, § 3). Поэтому область неопределенности имеет ширину $k2^{-i/3}$, где k порядка единицы. Соответственно линейная интерполяция перестает давать постоянное улучшение, как только мы достигли z , имеющего ошибку порядка $2^{-i/3}$. Однако если мы вычисляем определитель матрицы

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} - zI, \quad (40.3)$$

рассматривая ее как трехдиагональную, то область неопределенности, содержащая $z = 1$, порядка 2^{-i} , и корень при $z = 1$ может быть определен точно. (Об определении кратности см. § 57).

41. Этот пример может показаться слегка искусственным. К сожалению, трехдиагональная матрица с линейными элементарными делителями и собственным значением кратности r должна иметь $r - 1$ нулевых наддиагональных элементов; это следует просто из подсчета ранга. Однако если мы хотим рассмотреть собственные значения, которые в действительности не кратные, но равны с рабочей точностью, то мы можем дать значительно более интересные примеры. Хорошую иллюстрацию дает матрица W_{21}^+ из главы 5. Используя мантиссу с 30 двоичными знаками, доминирующий нуль был найден правильно до последней цифры, причем использовалась последовательная линейная интерполяция, несмотря на то, что второе собственное значение совпадает с первым с рабочей точностью. Рассматривались аналогичные примеры матриц, собственные значения которых с рабочей точностью были высокой кратности, и несмотря на то, что сходимость была чрезвычайно медленной, окончательная достижимая точность была очень высокой.

Другие интерполяционные методы

42. Аналогичные результаты можно установить для других интерполяционных методов, которые мы обсуждали. В каждом случае предельно точное значение z_k лежит в области неопределенности или ее непосредственной окрестности, и снова есть опасность неустойчивости после достижения предельной точности. Это явление было очень заметно для метода Мюллера, который мы использовали чаще, чем какой-либо другой интерполяционный метод. Наиболее часто случалось, что если итерации продолжались после достижения значения из области неопределенности, то получалась последовательность значений в основном с той же точностью, но рано или поздно происходил большой скачок. В экспе-

риментах этот скачок бывал настолько большим, что последующие итерации сходились к другому собственному значению! В библиотечных программах это исключалось при помощи ограничения текущей поправки (§ 62).

Один вопрос заслуживает специального упоминания. Можно подумать, что метод Мюллера не может точно локализовать двойное собственное значение, даже если метод вычисления таков, что область неопределенности очень мала. При этом можно рассуждать следующим образом.

Рассмотрим, например, полином $(z - 0,987623416)^2$. Если мы точно вычислим $f(z)$ в любых трех точках и выведем точный полином, проходящий через них, получим

$$z^2 - 1,975246832z + 0,975400011831509056. \quad (42.1)$$

Но если мы готовы использовать только девять десятичных знаков, мы не можем получить точно эти коэффициенты. В лучшем случае мы должны округлять в девятом знаке, и это ведет к ошибке в корнях порядка $(10^{-9})^{1/2}$. Следовательно, даже если мы вычислим $f(z)$ прямо из выражения $(z - 0,987623416)^2$, так что вычисленные значения будут иметь малую относительную ошибку, квадратичный полином, соответствующий этим значениям, должен дать сравнительно плохие результаты. В более практических ситуациях мы вряд ли будем получать значения функции в окрестности двойного нуля с точностью, достижимой в этом специальном случае, так что обычно ситуация будет еще более неприятной.

Неправильность этих аргументов состоит в том, что вычисления не проводятся таким образом. Если мы используем метод Мюллера (§ 20), мы получаем *поправку* для z_k на каждой итерации. Рассмотрим выражение разности $z_{k+1} - z_k$, полученное из (20.6). Имеем:

$$\begin{aligned} z_{k+1} - z_k &= \lambda_{k+1}(z_k - z_{k-1}) = \\ &= -2f_k \delta_k (z_k - z_{k-1}) / \{g_k \pm [g_k \pm 4f_k \delta_k \lambda_k (f_{k-2} \lambda_k - f_{k-1} \delta_k + f_k)]^{1/2}\}. \end{aligned} \quad (42.2)$$

Если бы вычисления были точными, величина в квадратных скобках была бы равна нулю для любых z_{k-2} , z_{k-1} , z_k . На практике это не так, но если максимальная относительная ошибка в f_k , f_{k-1} , f_{k-2} равна 2^{-s} , то знаменатель будет отличаться от верного значения самое большее в $s/2$ -ом знаке. Поэтому вычисленная *поправка* будет отличаться от поправки, соответствующей точным вычислениям, приблизительно так же. Следовательно, даже если точность вычисленных значений $f(z)$ уменьшается при приближении к собственному значению λ_p таким образом, что когда $|z_k - \lambda_p| = O(2^{-s})$, вычисленные значения имеют относительную ошибку 2^{-t+s} , последовательные итерации будут продолжать улучшаться на практике до тех пор, пока достигнутые значения не будут иметь относительную ошибку порядка 2^{-t} . Это в точности соответствует хорошо обусловленному двойному собственному значению в случае, когда методы вычислений дают область неопределенности диаметра порядка 2^{-t} .

Методы, в которых используются производные

43. Вообще говоря, методы, использующие одну или более производных, ведут себя совсем иначе, чем интерполяционные методы. Для изучения предельного поведения около простого собственного значения λ_p сделаем сначала естественное предположение, что в окрестности нуля относительная ошибка вычисленного значения производной меньше относительной ошибки вычисленного значения функции.

Для иллюстрации можно рассматривать метод Ньютона (§ 25), так как влияние ошибок округления на предельное поведение здесь в основном такое же, как и для всех таких методов. Если обозначить вычисленные значения $f(z_k)$ и $f'(z_k)$ через $\bar{f}_k$ и $\bar{f}'_k$ соответственно, то поправка, полученная на типичном шаге, равна $-\bar{f}_k/\bar{f}'_k$ вместо $-f_k/f'_k$. Можно написать

$$\bar{f}_k = f_k(1 + \varepsilon_k), \quad \bar{f}'_k = f'_k(1 + \eta_k), \quad (43.1)$$

и, по предположению, в окрестности λ_p имеем

$$|\eta_k| \ll |\varepsilon_k|. \quad (43.2)$$

Отношение между вычисленной поправкой и верной поправкой, следовательно, в основном определяется отношением $\bar{f}_k$ и f_k , и в обозначениях § 38 имеем

$$\bar{f}_k/f_k = \prod_{i=1}^n (\lambda'_i - z_k) / \prod_{i=1}^n (\lambda_i - z_k) \approx (\lambda'_p - z_k)/(\lambda_p - z_k). \quad (43.3)$$

До тех пор, пока $|\lambda_p - z_k|$ значительно больше $|\lambda_p - \lambda'_p|$, т. е. до тех пор, пока z_k достаточно далеко от области неопределенности, вычисленная поправка имеет малую относительную ошибку, но когда эти величины становятся одного порядка, это уже неверно. Если на этой стадии z_k есть $\lambda_p + O(\theta)$, где θ мало, то правильная поправка даст z_{k+1} вида $\lambda_p + O(\theta^2)$, причем мы предполагаем, что точность вычислений такова, что область неопределенности мала по сравнению с расстоянием от λ_p до соседних собственных значений. Следовательно, вычисленная поправка будет $\bar{f}_k[\lambda_p - z_k + O(\theta^2)]/f_k$, так что вычисленное z_{k+1} приблизительно равно λ'_p . Снова находим, что вычисленные итерации улучшаются до тех пор, пока не достигнута область неопределенности.

После этого поведение методов, использующих производные, отличается от поведения интерполяционных методов. Для точек z_k внутри области неопределенности вычисленные значения f' все равно будут иметь сравнительно малую относительную ошибку. Следовательно, отношение вычисленной поправки к верной поправке будет почти точно равно правой части (43.3), и так как верная поправка дает $z_{k+1} = \lambda_p$ почти точно, вычисленная поправка дает λ'_p почти точно. Следовательно, итерации будут колебаться более или менее случайно внутри области неопределенности.

44. Наше доказательство основано на предположении, что относительная ошибка вычисленных значений производной значительно меньше относительной ошибки вычисленных значений функции. Для кратных или очень близких собственных значений это будет, вообще говоря, верно для значений z вне области неопределенности, но для значений внутри области неопределенности для таких собственных значений это уже не будет справедливо. Более того вычисленное значение $f'(z)$ может даже быть нулем, а $f(z)$ отличным от нуля.

Однако замечания, сделанные в § 40, в полной мере приложимы к итерационным методам, в которых используются производные. Даже если матрица имеет кратный корень или очень близкие корни, но соответствующая область неопределенности мала, предельная точность будет велика; меняется только скорость сходимости. Более того, если мы случайно знаем кратность корня, то использование формул (25.9) или (28.12) даст высокую скорость сходимости.

Критерий достижения корня

45. Значительные трудности представляет вопрос о том, когда достигнутое итерационным процессом значение можно принять в качестве корня. Рассмотрим сначала упрощенную проблему, когда предполагается, что при вычислениях не делается ошибок округления, так что любая наперед заданная точность может быть достигнута при большом количестве шагов. Пусть вычисляется корень α . Для всех описанных методов в конечном счете

$$|z_{i+1} - \alpha| = O(|z_{i+1} - z_i|), \quad (45.1)$$

а для простых корней справедлив более сильный результат:

$$|z_{i+1} - \alpha| = o(|z_{i+1} - z_i|). \quad (45.2)$$

Следовательно, очевидным критерием будет

$$|z_{i+1} - z_i| < \varepsilon, \quad (45.3)$$

где ε — это некоторая предписанная «малая» величина. Заметим, однако, что если этот критерий должен быть применим для кратных корней, необходимо брать ε значительно меньшим максимальной ошибки, которая приемлема в определяемом корне. Например, если α — корень кратности r и используется метод Ньютона, то

$$\alpha \sim z_{i+1} + (r-1)(z_{i+1} - z_i), \quad (45.4)$$

и надо брать ε меньшим нашего предписанного значения ошибки по крайней мере в $(r-1)$ раз.

Заманчиво использовать какие-либо критерии, основанные на относительном изменении в текущей итерации, например

$$|(z_{i+1} - z_i)/z_i| < \varepsilon. \quad (45.5)$$

Для сходимости к корню $z = \alpha \neq 0$, конечно, такой критерий хорош, но если $\alpha = 0$, то

$$|(z_{i+1} - z_i)/z_i| \rightarrow 1 \quad (45.6)$$

для простого корня. Если мы имеем дело с явным полиномом, то нулевые собственные значения легко определяются и отделяются. Следовательно, в этом случае, если все ненулевые корни простые, может быть использован соответствующий критерий.

Влияние ошибок округления

46. Нет особого смысла продолжать анализ предыдущего параграфа, так как в действительности ошибки округления — основной фактор при практическом решении задачи. На практике возможная достижимая точность вычислений ограничивается областью неопределенности корня, и, вообще говоря, у нас нет а priori информации о размере этой области. Если мы в критерии (45.3) выберем слишком малое значение ε , то он может не быть удовлетворенным, как бы долго ни продолжали мы итерации. С другой стороны, если ε будет слишком велико, то достигнутая точность будет значительно меньше точности, которую можно было бы получить. Это будет верно, например, в случае *хорошо определенных кратных* корней, так как для них сходимость к пределу медленная, но неуклонная.

Если область неопределенности, окружающая корень, сравнительно велика, то может оказаться невозможным удовлетворить самым скромным требованиям. Например, если используется форма Фробениуса и матрица имеет собственные значения 1, 2, . . . , 20, то при работе с обычной точностью на АСЕ (почти 14 десятичных знаков) для сходимости к корням 10, 11, . . . , 19 нельзя взять значение ε , даже равное 10^{-2} . На самом деле область неопределенности этих собственных значений настолько велика, что невозможно определить, когда итерации попадут в окрестность собственных значений, если не работать со значительно более высокой точностью, чем 14 десятичных знаков.

Идеальная программа должна была бы быть способной определять случаи, когда необходима более высокая точность вычислений, и отличать их от неупорядоченного продвижения, предшествующего локализации корня. Кроме того, она должна прогрессивно увеличивать точность вычислений для получения собственных значений с предписанной точностью. На практике обычно цели значительно более скромные, и используется критерий (45.3) с некоторым компромиссным значением ε .

47. Гарвик предложил простой прием, уменьшающий опасность бесконечности числа итераций для плохо обусловленных корней без уменьшения достижимой точности для хорошо обусловленных корней. Этот прием основан на том, что при движении по направлению к корню, начиная с некоторой стадии, разности $|z_{i+1} - z_i|$ постоянно уменьшаются, до тех пор, пока не достигается область неопределенности. После этого эти разности могут вести себя неустойчивым образом, который мы описали. Гарвик предлагает использовать сравнительно простой критерий вида

$$|z_{i+1} - z_i| < \varepsilon, \quad (47.1)$$

а после удовлетворения критерия продолжать итерации до тех пор, пока разности между последующими z_i уменьшаются. Когда же, наконец, будут достигнуты значения, для которых

$$|z_{r+2} - z_{r+1}| \geq |z_{r+1} - z_r|, \quad (47.2)$$

то предлагается считать значение z_{r+1} приемлемым.

Для иллюстрации использовался метод Ньютона для явного полинома с нулями 1, 2, . . . , 20, причем вычисления производились на АСЕ с двойной точностью (около 28 десятичных знаков) при $\varepsilon = 10^{-6}$. Было обнаружено, что для каждого корня итерации продолжались до достижения значений внутри области неопределенности. Так, ошибка в полученном приближении для корня $z = 1$ была порядка 10^{-27} , а для плохо обусловленного корня $z = 15$ порядка 10^{-14} , так что была достигнута максимальная эффективность.

Кажется, мало что сделано для получения программ, автоматически определяющих, когда достигнута предельная точность. На вычислительных машинах, использующих ненормализованную арифметику с плавающей запятой, это не должно быть слишком трудным, так как из предыдущих рассмотрений очевидно, что итерации должны продолжаться до тех пор, пока текущие вычисленные значения функции имеют несколько верных значащих цифр. Как только мы достигли значения z_i , для которого это неверно, z_i не отличимо от корня функции при данной точности вычислений и использованном алгоритме.

Удаление вычисленных корней

48. После того как корень $f(z)$ уже получен, нужно избежать того, чтобы в следующих итерациях получать сходимость к этому корню. Если мы работаем с явным полиномом, то после получения корня α можно точно вычислить полином $f(z)/(z - \alpha)$. Аналогично, если был получен квадратичный множитель $z^2 - pz - l$, мы можем вычислить $f(z)/(z^2 - pz - l)$. Этот процесс обычно называется процессом *исчерпывания*. Заметим из (2.10), что

$$f(z)/(z - \alpha) = \sum_{r=1}^n s_r(\alpha) z^{r-1}, \quad (48.1)$$

где $s_r(\alpha)$ — величины из схемы Горнера при $z = \alpha$, а в обозначениях § 31

$$f(z)/(z^2 - pz - l) = \sum_{r=2}^n q_r z^{r-2}. \quad (48.2)$$

На практике полученные корень или квадратичный множитель обычно имеют ошибки, и в процессе исчерпывания возникают дополнительные ошибки. Следовательно, есть опасность, что точность последовательно вычисленных корней будет постоянно уменьшаться, и естественно, что процесс исчерпывания обычно признается чрезвычайно опасным в этом отношении.

В главе 2 работы Уилкинсона (1963b) дан очень детальный анализ этой проблемы, и здесь мы ограничимся изложением главных результатов.

Если перед каждым исчерпыванием *итерации продолжались до получения предельной точности для данной точности вычислений*, то при условии, что корни вычисляются примерно в порядке увеличения абсолютной величины, происходит только малое последовательное ухудшение. Вообще говоря, точность вычисления корня при итерациях полинома после исчерпывания очень мало уменьшается по сравнению с точностью вычисления корня исходного полинома. С другой стороны, если корни полинома сильно отличаются по абсолютной величине и найдены приближения для одного из больших корней, то проведение исчерпывания перед нахождением значительно меньших нулей может привести к катастрофической потере точности. *Это, вероятно, верно, даже если все корни исходного полинома очень хорошо обусловлены, в том смысле, что они не чувствительны к малым изменениям коэффициентов.*

Хотя опасность исчерпывания часто преувеличивают, кажется, нет надежных методов, которые обеспечивают в общем случае нахождение корней примерно в порядке возрастания абсолютной величины, и в § 55 мы опишем методы для удаления вычисленных нулей без проведения явного исчерпывания.

Исчерпывание для матриц Хессенберга

49. Для явного полинома после определения корня можно определить явный полином, имеющий остальные $(n - 1)$ корней. Естественно спросить, существуют ли аналогичные методы для каждой из остальных специальных форм, и если есть, то будут ли они устойчивы. Вопрос можно точно поставить следующим образом.

Если известно собственное значение матрицы Хессенберга (трехдиагональной) порядка n , можно ли получить матрицу Хессенберга (трех-

диагональную) порядка $(n - 1)$, имеющую остальные $(n - 1)$ собственных значений?

Рассмотрим сначала случай матрицы Хессенберга. Пусть λ_1 — собственное значение верхней матрицы Хессенберга A . Как всегда, можно предположить, что все поддиагональные элементы A отличны от нуля. Если в методе Химана, описанном в § 11, взять $z = \lambda_1$, то определится матрица M вида

$$M = \begin{bmatrix} 1 & & -x_1 \\ & 1 & -x_2 \\ & & \ddots \\ & & & -x_{n-1} \\ & & & & 1 \end{bmatrix}, \quad (49.1)$$

такая, что элементы последнего столбца $B = (A - \lambda_1 I) M$ равны нулю, кроме первого, который пропорционален $\det(A - \lambda_1 I)$. Так как λ_1 есть собственное значение A , $\det(A - \lambda_1 I)$ равен нулю, и, следовательно, весь последний столбец равен нулю. Заметим, что с точностью до последнего столбца B совпадает с $(A - \lambda_1 I)$, так что при $n = 5$ B имеет вид

$$B = \begin{bmatrix} \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & 0 \\ & \times & \times & \times & 0 \\ & & \times & \times & 0 \\ & & & \times & 0 \end{bmatrix}. \quad (49.2)$$

Можно завершить преобразование подобия, умножив B слева на M^{-1} , что соответствует прибавлению n -й строки, умноженной на x_i , к i -й строке для всех i от 1 до $n - 1$. Ясно, что это не нарушит формы Хессенберга и нулей в последнем столбце, так что $M^{-1}(A - \lambda_1 I)M$ также имеет вид (49.2). Следовательно, $M^{-1}(A - \lambda_1 I)M + \lambda_1 I$ имеет вид

$$\left[\begin{array}{cccc|c} \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & 0 \\ & \times & \times & \times & 0 \\ & & \times & \times & 0 \\ \hline & & & \times & \lambda_1 \end{array} \right]. \quad (49.3)$$

Таким образом, оставшиеся собственные значения равны собственным значениям матрицы Хессенберга порядка $(n - 1)$ в верхнем левом углу.

50. Так как трехдиагональные матрицы являются частным случаем матриц Хессенберга, к ним можно приложить описанный выше процесс. Умножение справа на M не меняет первые $(n - 1)$ столбцов, и форма (49.2) имеет вид

$$\begin{bmatrix} \times & \times & & & 0 \\ \times & \times & \times & & 0 \\ & \times & \times & \times & 0 \\ & & \times & \times & 0 \\ & & & \times & 0 \end{bmatrix}. \quad (50.1)$$

Умножение слева на M^{-1} меняет только $(n - 1)$ -й столбец так, что

окончательная матрица имеет вид

$$\begin{bmatrix} \times & \times & & \times & 0 \\ \times & \times & \times & \times & 0 \\ & \times & \times & \times & 0 \\ & & \times & \times & 0 \\ & & & \times & 0 \end{bmatrix}. \quad (50.2)$$

Матрица порядка $(n - 1)$ в верхнем левом углу уже не трехдиагональная; она имеет дополнительные ненулевые элементы в целом последнем столбце. Следовательно, трехдиагональная форма не инвариантна по отношению к нашему процессу исчерпывания.

Принято называть матрицы, полученные из трехдиагональных добавлением ненулевых элементов в последнем столбце, *дополненными трехдиагональными матрицами*. Из нашего рассмотрения сразу следует, что дополненные трехдиагональные матрицы *инвариантны* по отношению к нашему процессу исчерпывания. Более точно, процесс исчерпывания матрицы A_n приводит к матрице C_n , причем A_n и C_n имеют вид

$$A_n = \begin{bmatrix} \times & \times & & & & \times \\ \times & \times & \times & & & \times \\ & \times & \times & \times & & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times & \times \end{bmatrix},$$

$$C_n = \begin{bmatrix} \times & \times & & & \times & 0 \\ \times & \times & \times & & \times & 0 \\ & \times & \times & \times & \times & 0 \\ & & \times & \times & \times & 0 \\ & & & \times & \times & \times & 0 \\ & & & & \times & \times & 0 \\ & & & & & \times & \times & 0 \\ \hline & & & & & & \times & \lambda_1 \end{bmatrix}. \quad (50.3)$$

Таким образом, хотя при исчерпывании трехдиагональная форма и не сохраняется, матрица при этом не становится значительно сложнее.

Исчерпывание для трехдиагональной матрицы

51. Процесс исчерпывания для трехдиагональной матрицы был выведен как специальный случай процесса для *сержных* матриц Хессенберга, но этот процесс имеет тесную связь с процессом §8, если ненулевые элементы вводить в первом столбце матрицы после исчерпывания вместо последнего. Проиллюстрируем это для матрицы пятого порядка. Предположим, что наша исходная трехдиагональная матрица A_n имеет вид

$$A_n = \begin{bmatrix} \alpha_1 & 1 & & & \\ \beta_2 & \alpha_2 & 1 & & \\ & \beta_3 & \alpha_3 & 1 & \\ & & \beta_4 & \alpha_4 & 1 \\ & & & \beta_5 & \alpha_5 \end{bmatrix}. \quad (51.1)$$

Если λ_1 — собственное значение A_n , определим $p_r(\lambda_1)$ соотношениями $p_0(\lambda_1) = 1$, $p_1(\lambda_1) = \alpha_1 - \lambda_1$, $p_r(\lambda_1) = (\alpha_r - \lambda_1)p_{r-1}(\lambda_1) - \beta_r p_{r-2}(\lambda_1)$ ($r = 2, \dots, n$). (51.2)

Пусть M — матрица вида

$$M = \begin{bmatrix} 1 & & & & \\ -p_1(\lambda_1) & 1 & & & \\ p_2(\lambda_1) & & 1 & & \\ -p_3(\lambda_1) & & & 1 & \\ p_4(\lambda_1) & & & & 1 \end{bmatrix}. \quad (51.3)$$

Тогда

$$(A_n - \lambda_1 I)M = \begin{bmatrix} 0 & 1 & & & \\ 0 & (\alpha_2 - \lambda_1) & 1 & & \\ 0 & \beta_3 & (\alpha_3 - \lambda_1) & 1 & \\ 0 & & \beta_4 & (\alpha_4 - \lambda_1) & 1 \\ 0 & & & \beta_5 & (\alpha_5 - \lambda_1) \end{bmatrix}, \quad (51.4)$$

причем $(n, 1)$ -й элемент равен нулю, так как $p_n(\lambda_1) = 0$. Следовательно, мы имеем

$$M^{-1}(A_n - \lambda_1 I)M + \lambda_1 I = \begin{bmatrix} \lambda_1 & 1 & & & \\ 0 & \alpha_2 + p_1(\lambda_1) & 1 & & \\ 0 & \beta_3 - p_2(\lambda_1) & \alpha_3 & 1 & \\ 0 & p_3(\lambda_1) & \beta_4 & \alpha_4 & 1 \\ 0 & -p_4(\lambda_1) & & \beta_5 & \alpha_5 \end{bmatrix}. \quad (51.5)$$

Таким образом, после исчерпывания матрица A_{n-1} порядка $(n-1)$ в нижнем правом углу имеет дополнительные элементы $\pm p_r(\lambda_1)$ в первом столбце.

Обозначим через $q_{r+1}(z)$ главные ведущие миноры $(A_{n-1} - zI)$ порядка r . Имеем

$$\left. \begin{aligned} q_1(z) &= 1, & q_2(z) &= (\alpha_2 - z)q_1(z) + p_1(\lambda_1), \\ q_r(z) &= (\alpha_r - z)q_{r-1}(z) - \beta_r q_{r-2}(z) + p_{r-1}(\lambda_1) \end{aligned} \right\} \quad (r = 3, \dots, n), \quad (51.6)$$

откуда видно, что добавленный столбец элементов $p_r(\lambda_1)$ вряд ли усложняет рекуррентные соотношения.

Для выполнения второго исчерпывания, после того как определено собственное значение λ_2 матрицы A_{n-1} , умножаем справа на матрицу вида

$$\begin{bmatrix} 1 & & & \\ -q_2(\lambda_2) & 1 & & \\ q_3(\lambda_2) & & 1 & \\ -q_4(\lambda_2) & & & 1 \end{bmatrix}, \quad (51.7)$$

так что окончательная матрица порядка $(n-2)$ имеет вид

$$\begin{bmatrix} \alpha_3 + q_2(\lambda_2) & 1 & & \\ -q_3(\lambda_2) & \alpha_4 & 1 & \\ q_4(\lambda_2) & \beta_5 & \alpha_5 & \end{bmatrix}. \quad (51.8)$$

Следовательно, рекуррентные соотношения, связывающие главные миноры матриц после исчерпывания, не становятся сложнее соотношений (51.6).

Исчерпывание при помощи вращений или устойчивых элементарных преобразований

52. Процесс исчерпывания для матриц Хессенберга формально тесно связан с методом Химана, а мы видели, что последний исключительно устойчив. Исчерпывание трехдиагональных матриц связано с простыми рекуррентными соотношениями, использующимися в § 8, и они также очень устойчивы.

Тем не менее процессы исчерпывания не обладают той же численной устойчивостью. Как можно понять, это происходит потому, что не используются перестановки. Но если они введены или если используются плоские вращения, то можно получить устойчивые формы исчерпывания. Рассмотрим их.

В § 14 было рассмотрено исключение последних $(n - 1)$ элементов последнего столбца матрицы Хессенберга при помощи последовательного умножения справа на вращения в плоскостях $(n - 1, n)$, $(n - 2, n)$, $\dots$, $(1, n)$. В частном случае, когда $z = \lambda_1$, собственному значению A , первый элемент преобразованной матрицы также должен быть равен нулю (если вычисления точные), и, следовательно, вычисленная матрица будет иметь такой же вид, как в (49.2). Преобразование подобия завершается умножением слева на вращения в плоскостях $(n - 1, n)$, $(n - 2, n)$, $\dots$, $(1, n)$. Очевидно, что каждое вращение сохраняет нули в последнем столбце и хессенбергову природу матрицы в верхнем левом углу, но вращение в плоскости (r, n) ($r = n - 1, \dots, 2$) вводит отличный от нуля элемент в позиции $(n, r - 1)$. Окончательная матрица, полученная после прибавления $\lambda_1 I$, имеет вид

$$\left[\begin{array}{cccc|c} \times & \times & \times & \times & 0 \\ \times & \times & \times & \times & 0 \\ & & \times & \times & 0 \\ & & & \times & 0 \\ \hline \times & \times & \times & \times & \lambda_1 \end{array} \right]. \quad (52.1)$$

Присутствие ненулевых элементов в последней строке неважно, и матрица Хессенберга порядка $(n - 1)$ в верхнем левом углу снова имеет $(n - 1)$ оставшееся собственное значение. Следовательно, форма Хессенберга инвариантна по отношению к нашему процессу исчерпывания, даже если используются плоские вращения. В точности эквивалентный процесс исчерпывания может быть выполнен с использованием устойчивых неунитарных элементарных преобразований, и он также приводит к матрице вида (52.1).

Прежде чем обсуждать влияние ошибок округления, рассмотрим те же процессы, примененные к трехдиагональной матрице. Ясно, что после окончания умножений справа преобразованная матрица имеет вид (50.1).¹

Покажем, что умножение слева приводит к матрице вида (52.1), т. е. трехдиагональная матрица заменяется, вообще говоря, полной матрицей Хессенберга. Доказательство по индукции. Предположим, что после умножения слева на вращение в плоскости (r, n) , матрица, например, для

случая $n = 7$, $r = 4$ имеет вид

$$\left[\begin{array}{cccccc|c} \times & \times & & & & & 0 \\ & \times & \times & & & & 0 \\ & & \times & \times & \times & & 0 \\ & & & \times & \times & \times & \times & 0 \\ & & & & \times & \times & \times & 0 \\ & & & & & \times & \times & 0 \\ 0 & 0 & \times & \times & \times & \times & & 0 \end{array} \right]. \quad (52.2)$$

При следующем вращении в плоскости $(r-1, n)$ строки $r-1$ и n заменяются своими линейными комбинациями. Следовательно, вообще говоря, $(r-1)$ -я строка получает ненулевые элементы в позициях от $r+1$ до $n-1$, а n -я строка — в позиции $r-2$. Снова элементарные устойчивые неунитарные преобразования дадут такой же эффект, хотя соответствующая матрица Хессенберга будет иметь большое количество нулевых элементов. Однако если исходная трехдиагональная матрица была симметричной, ортогональное исчерпывание сохраняет симметрию, и можно показать поэтому, что сохраняется трехдиагональный вид. Доказательство следующее.

Так как последний столбец окончательной матрицы равен нулю, должна равняться нулю и последняя строка. Это в свою очередь означает, что после умножения слева на вращение в плоскости (r, n) элементы в позициях $(n, r+1)$, $(n, r+2)$, ..., $(n, n-1)$ уже должны быть равны нулю. Действительно, единственная операция, которую осталось выполнить, это последовательная замена строки n на

$$(\text{строка } n) \cos \theta_i - (\text{строка } i) \sin \theta_i \quad (i = r-1, \dots, 1) \quad (52.3)$$

(здесь используются очевидные обозначения), и из того, что поддиагональные элементы исходной матрицы не равны нулю, следует, что не может быть $\cos \theta_i = 0$. Так как согласно (52.2) элементы в позициях $r+1, \dots, (n-1)$ каждой строки $r-1, \dots, 1$ нулевые, соответствующие элементы n -й строки уже должны быть равны нулю. Отсюда следует, что не вводится ненулевых элементов над диагональю, так что окончательная матрица трехдиагональная. Исчерпывание, использующее устойчивые неунитарные матрицы, не сохраняет симметрии или трехдиагонального вида.

Устойчивость исчерпывания

53. Рассмотрим влияние ошибок округления в самом процессе исчерпывания и в значении λ_1 . Сначала обсудим исчерпывание матрицы Хессенберга с использованием плоских вращений.

Если рассматриваются умножения справа, наш общий анализ гл. 3, §§ 20—26 дает, что если $R_{n-1, n}$, $R_{n-2, n}$, ..., R_{1n} суть точные плоские вращения, соответствующие последовательным вычисленным матрицам, то окончательная вычисленная матрица будет очень близка к $AR_{n-1, n}R_{n-2, n} \dots R_{1n}$. Этот анализ оправдывает введение нулей в позициях (n, n) , $(n-1, n)$, ..., $(2, n)$. Возникновение нуля в позиции $(1, n)$, которое гарантируется в случае точных вычислений, не является, конечно, прямым следствием специального вращения, и наш анализ не показывает, что этот элемент должен быть мал, даже если λ_1 верно с рабочей точностью. Легко построить матрицы, проблема собственных

значений которых хорошо обусловлена, но вычисленный $(1, n)$ -элемент после завершения умножения справа весьма велик, и следовательно, окончательная вычисленная матрица $R_{1n}^T R_{2n}^T \dots R_{n-1, n}^T (A - \lambda_1 I) R_{n-1, n} \dots R_{2n} R_{1n}$ имеет вид

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & 0 \\ & & \times & \times & \times & 0 \\ & & & \times & \times & 0 \\ & & & & \times & 0 \\ \hline \times & \times & \times & \times & \times & \times \end{bmatrix}, \quad (53.1)$$

причем ни элемент в позиции $(1, n)$, ни элемент в позиции (n, n) не малы.

Аналогично при использовании плоских вращений для исчерпывания симметричной трехдиагональной матрицы окончательная вычисленная матрица может иметь вид

$$\begin{bmatrix} \times & \times & & & & \times \\ \times & \times & \times & & & 0 \\ & \times & \times & \times & & 0 \\ & & \times & \times & \times & 0 \\ & & & \times & \times & 0 \\ \hline \times & 0 & 0 & 0 & 0 & \times \end{bmatrix}, \quad (53.2)$$

причем три подчеркнутых элемента, которые должны бы быть равными нулю при точных вычислениях, могут сильно отличаться от нуля.

Большинство из этих замечаний приложимо к процессу исчерпывания, использующему элементарные устойчивые неунитарные преобразования. Покажем, что для таких процессов исчерпывания срыв аннулирования $(1, n)$ -элемента после завершения умножения справа тесно связан с явлением, которое обсуждалось в гл. 5 § 56.

Рассмотрим матрицу A порядка 21 вида

$$\begin{bmatrix} -10 & 1 & & & & \\ & 1 & -9 & 1 & & \\ & & 1 & -8 & & \\ & & & \ddots & & \\ & & & & 1 & 1 \\ & & & & & 1 & 9 & 1 \\ & & & & & & 1 & 10 \end{bmatrix}, \quad (53.3)$$

собственные значения которой совпадают с собственными значениями W_{21}^- . Собственное значение $\lambda_1 = 10,7461942$ верно с 9 десятичными значащими цифрами. Давайте выполним процесс исчерпывания для $(A - \lambda_1 I)$, используя элементарные устойчивые неунитарные преобразования.

Сначала сравниваем абсолютные значения элементов в позициях $(21, 20)$ и $(21, 21)$, и если модуль элемента в позиции $(21, 21)$ больше, переставляем столбцы 21 и 20. Затем вычитаем кратное 20-го столбца из 21-го так, чтобы получить ноль в позиции $(21, 21)$. Сразу видно, что эти действия в точности совпадают с действиями в первом шаге метода Гаусса с перестановками для $(W_{21}^- - \lambda_1 I)$. Это совпадение двух процессов сохраняется

все время. Заметим, однако, что наличие перестановки на какой-либо стадии метода Гаусса соответствует отсутствию перестановки в процессе исчерпывания, и наоборот. После завершения умножения справа столбцы 20, 19, . . . , 1 совпадают с ведущими строками 1, 2, . . . , 20, полученными в методе Гаусса, а 21-й столбец приведенной $(A - \lambda_1 I)$ равен последней ведущей строке, полученной в методе Гаусса, так что элемент в позиции (1, 21), который должен бы быть равен нулю, на самом деле равен $-20,6954139$. Элементы столбца $(21 - i)$ суть w_i, v_i, u_i из табл. 10 гл. 5.

Общие замечания об исчерпывании

54. В главе 8 будут описаны несколько методов исчерпывания, которые и экономичны, и устойчивы, но так как они тесно связаны с обсуждаемыми там вопросами, мы сейчас отложим их рассмотрение.

Все приемы исчерпывания, описанные в этой главе, в том или другом смысле не удовлетворительны. Вообще с целью сохранения собственных значений исчерпанной матрицы они вынуждают вычислять собственные значения с излишне высокой точностью и использовать высокую точность вычислений. Этого можно избежать в случае явного полинома, если собственные значения могут быть найдены в порядке возрастания абсолютной величины, но, вообще говоря, нет простого метода, который бы это гарантировал. Поэтому естественно поставить вопрос, существуют ли другие методы удаления вычисленных нулей без явного исчерпывания.

Удаление вычисленных нулей

55. Пусть $f(z)$ — некоторый полином, заданный не обязательно явно, и пусть $\lambda_1, \lambda_2, \dots, \lambda_n$ — его корни. Если получены приближения $\lambda'_1, \lambda'_2, \dots, \lambda'_r$ его первых r корней, то можно определить функцию

$$g_r(z) \equiv f(z) / \prod_{i=1}^r (z - \lambda'_i). \quad (55.1)$$

Если бы каждый λ'_i был в точности равен соответствующему λ_i , то $g_r(z)$ была бы полиномом степени $n - r$, имеющим остальные корни $\lambda_{r+1}, \dots, \lambda_n$. Однако, даже если λ'_i неточны, $g_r(z)$ будет, вообще говоря, иметь корни $\lambda_{r+1}, \dots, \lambda_n$. Действительно, $g_r(z)$ имеет эти корни независимо от λ'_i , если только эти величины случайно не совпадут с некоторыми из $\lambda_{r+1}, \dots, \lambda_n$. Предполагая, что можно вычислить значение $f(z)$ при предписанном значении z , очевидно, можно вычислить $g_r(z)$, просто поделив вычисленное значение $f(z)$ на вычисленное значение $\prod_{i=1}^r (z - \lambda'_i)$. Следовательно, для определения корней $g_r(z)$ можно сразу приложить методы, основанные лишь на значениях функции.

Для приложения метода Ньютона необходимо определение $g_r(z)/g'_r(z)$, но так как

$$g'_r(z)/g_r(z) = f'(z)/f(z) - \sum_{i=1}^r 1/(z - \lambda'_i), \quad (55.2)$$

то эту функцию мы сможем вычислить, если мы можем вычислять $f(z)$ и $f'(z)$. И наконец, для методов, использующих $g''_r(z)$, имеем

$$[g'_r(z)/g_r(z)]^2 - g''_r(z)/g_r(z) = [f'(z)/f(z)]^2 - f''(z)/f(z) - \sum_{i=1}^r 1/(z - \lambda'_i)^2. \quad (55.3)$$

Таким образом, все, что требуется для итерационных методов, вычисляется.

Удаление вычисленного квадратичного множителя

56. Существуют аналогичные приемы для удаления полученных квадратичных множителей, которые могут использоваться в связи с методами типа Берстоу. Предположим, что получен квадратичный множитель $z^2 - Pz - L$. Если он является точным множителем и можно вычислить явный полином $g(z) = f(z)/(z^2 - Pz - L)$, то существенным требованием в шаге метода Берстоу является вычисление остатков при делении $g(z)$ последовательно два раза на трехчлен $z^2 - pz - l$. Поэтому нужны формулы, которые дают эти остатки при работе прямо с $f(z)$. Предположим, что

$$f(z) \equiv T(z)(z^2 - pz - l)^2 + (cz + d)(z^2 - pz - l) + az + b, \quad (56.1)$$

так что a, b, c, d это величины q_1, q'_0, T_1, T'_0 §§ 31, 32. Если

$$g(z) \equiv V(z)(z^2 - pz - l)^2 + (c''z + d')(z^2 - pz - l) + a''z + b'', \quad (56.2)$$

то

$$\begin{aligned} (z^2 - Pz - L)[V(z)(z^2 - pz - l)^2 + (c''z + d')(z^2 - pz - l) + a''z + b''] &\equiv \\ &\equiv T(z)(z^2 - pz - l)^2 + (cz + d)(z^2 - pz - l) + az + b. \end{aligned} \quad (56.3)$$

Это тождество позволяет выразить a'', b'', c'', d'' через a, b, c, d и, следовательно, получить остатки, соответствующие $g(z)$, по остаткам, соответствующим $f(z)$.

Следуя Хандскомбу (1962), но слегка изменив обозначения, можно решение выразить в виде

$$\left. \begin{aligned} p' &= p - P, & l' &= l - L, \\ f &= pp' + l, & e &= fl' - lp'^2, \\ a' &= al' - bp', & b' &= bf - alp', \\ c' &= ce - a', & d' &= de - b' - ap', \\ a'e &= a', & b'e &= b', \\ c'e^2 &= c'l' - d'p', & d'e^2 &= d'f - c'lp'. \end{aligned} \right\} \quad (56.4)$$

Заметим, что если a и b стремятся к нулю, то это же верно для a'' и b'' . Более того, так как сходимость для a и b квадратичная, это же верно и для a'' и b'' . Поэтому получается хорошая сходимость к корням $f(z)$, даже если начальный множитель $z^2 - Pz - L$ был очень плохой. Так как нам нужны a, b, c и d с точностью до произвольного множителя, мы можем взять

$$\left. \begin{aligned} a'' &= a'e, & b'' &= b'e, \\ c'' &= c'l' - d'p', & d'' &= d'f - c'lp'. \end{aligned} \right\} \quad (56.5)$$

Очевидно, что, начиная с a'', b'', c'', d'' , можно удалить этим же методом второй квадратичный множитель и т. д.

Общие комментарии о методах удаления

57. На первый взгляд эти процессы кажутся опасными, так как функция $g_r(z)$ будет, вообще говоря, иметь своими корнями все корни $f(z)$ и полюсы в точках λ_i ($i = 1, \dots, r$), причем эти полюсы будут очень близки к соответствующим корням. Однако на практике работают с вычисленными значениями $g_r(z)$, а они в свою очередь получают по вычисленным значениям $f(z)$. Существенной чертой практического процесса

является то, что можно использовать все время одну и ту же точность вычислений.

Рассмотрим вычисление типичного значения $g_r(z)$. Вычисленное значение $f(z)$ равно $\prod_{i=1}^n (z - \bar{\lambda}_i)$, где $\bar{\lambda}_i$ — значения из областей неопределенности соответствующих λ_i , которые являются, конечно, функциями z . Принятые значения $\bar{\lambda}_i$ также будут из областей неопределенности, или очень близки к ним, и, следовательно, вычисленное значение $g_r(z)$ равно $\prod_{i=1}^n (z - \bar{\lambda}_i) / \prod_{i=1}^r (z - \lambda'_i)$. Для значений z , не слишком близких к какому-либо λ_i ($i = 1, \dots, r$), вычисленное значение $g_r(z)$ поэтому в основном равно $\prod_{i=r+1}^n (z - \bar{\lambda}_i)$. Следовательно, можно ожидать, что любой из оставшихся корней λ_i ($i = r + 1, \dots, n$) будет вычислен столь же точно с использованием значений $g_r(z)$, как и с использованием $f(z)$. Если z из окрестности какого-либо λ_i ($i = 1, \dots, r$), то вычисленные значения будут настолько меняться, что вопроса о сходимости снова к одному из этих нулей не возникает. Естественно, мы должны пресекать попытки вычисления в любой точке $z = \lambda'_i$, и это нетрудно сделать.

В этих рассуждениях предполагалось, что λ_i — простые корни. Однако кратность или патологическая близость корней не вызывает никаких специальных затруднений. Они определяются с меньшей точностью только тогда, когда велики области неопределенности, и это будет так для явных полиномов. Так, если $f(z)$ имеет двойной корень при $z = \lambda_1 = \lambda_2$, а принятое значение корня равно λ'_1 , вычисленные значения $f(z)/(z - \lambda'_1)$ ведут себя так, как если бы функция имела простой корень в окрестности $z = \lambda_1$, при условии, что z находятся вдали от области неопределенности. На АСЕ вычислялись корни кратностей до пяти с использованием этого метода удаления, и все они были получены с ошибками, не большими тех, которые можно было бы ожидать из рассмотрения размеров областей неопределенности.

Заметим, что случайный выбор в качестве корня значения с серьезной ошибкой не влияет на точность вычисления остальных корней, но если принято значение λ'_i , не имеющее предельной точности, то вновь вероятна сходимость к верному предельному приближению λ_i . Так как программы обычно предназначены для определения n нулей полинома степени n , выбор ложного корня обычно ведет к потере настоящего.

Важной чертой методов удаления является то, что каждый корень вычисляется независимо. Поэтому совпадение суммы вычисленных корней со следом исходной матрицы является хорошей проверкой их точности. Это неверно для большинства методов исчерпывания. Их природа часто такова, что сумма вычисленных нулей верна, даже если корни мало связаны с собственными значениями. Например, для исчерпывания явного полинома, даже если мы на каждом шагу, кроме последнего (на этой стадии полином линейный, и корень определяется без итераций!), примем в качестве корня случайную величину, легко видеть, что сумма будет верна.

Асимптотическая скорость сходимости

58. То, что метод Ньютона (§ 25) и метод Лагерра (§ 28) остаются квадратично и кубически сходящимися при применении к $g_r(z)$, не очевидно непосредственно. Покажем, что это действительно верно. Рассмотрим

сначала применение метода Ньютона и положим $z = \lambda_{r+1} + h$. Тогда

$$g'_r(z)/g_r(z) = \sum_{i=1}^n 1/(z - \lambda_i) - \sum_{i=1}^r 1/(z - \lambda_i) = 1/h + \sum_{i=1}^n 1/(z - \lambda_i) - \sum_{i=1}^r 1/(z - \lambda_i), \quad (58.1)$$

где $\sum'$ обозначает, что $i = r + 1$ пропускается в сумме. Отсюда

$$g'_r(z)/g_r(z) = 1/h + A + O(h), \quad (58.2)$$

где

$$A = \sum'_{i=1}^n 1/(\lambda_{r+1} - \lambda_i) - \sum_{i=1}^r 1/(\lambda_{r+1} - \lambda_i), \quad (58.3)$$

что дает

$$z - g_r(z)/g'_r(z) = \lambda_{r+1} + Ah^2 + O(h^3). \quad (58.4)$$

При методе Лагерра мы пользуемся формулой

$$z_{k+1} = z_k - \frac{(n-r)g_r(z_k)}{g'_r(z_k) \pm \{(n-r-1)^2 [g'_r(z_k)]^2 - (n-r)(n-r-1)g_r(z_k)g''_r(z_k)\}^{1/2}}, \quad (58.5)$$

и, снова положив $z_k = \lambda_{r+1} + h$ и воспользовавшись (55.2) и (55.3), можно проверить, что

$$z_{k+1} - \lambda_{r+1} \sim \frac{1}{2} h^3 [(n-r-1)B - A^2]/(n-r-1), \quad (58.6)$$

где A определяется так же, как в (58.3), а B определяется формулой

$$B = \sum'_{i=1}^n 1/(\lambda_{r+1} - \lambda_i)^2 - \sum_{i=1}^r 1/(\lambda_{r+1} - \lambda_i)^2. \quad (58.7)$$

Сходимость в целом

59. До сих пор мы не касались вопроса построения начальных приближений к собственным значениям. Разумеется, если мы располагаем приближениями из соображений, вытекающих из смысла задачи, методы этой главы вполне удовлетворительны, но практически они часто используются как единственное средство локализации собственных значений.

Если известно, что все собственные значения вещественны, то ситуация вполне удовлетворительна. В отношении метода Лагерра мы знаем, что сходимость к соседним корням обеспечена при любом начальном приближении. Простейший способ — использовать $\|A\|_\infty$ в качестве исходного значения каждого корня и применить метод удаления, описанный в § 55. Сходимость в целом столь замечательна, что этот простой процесс вполне удовлетворителен, если только A не имеет кратных или патологически близких корней. Для таких корней (28.12) дает, что

$$z_{k+1} - \lambda_m \sim C(z_k - \lambda_m) \quad (59.1)$$

и

$$z_{k+2} - \lambda_m \sim C(z_{k+1} - \lambda_m), \quad (59.2)$$

откуда

$$(z_{k+2} - z_{k+1}) \sim C (z_{k+1} - z_k). \quad (59.3)$$

Положив $z_{k+1} - z_k = \Delta z_k$, получим

$$\Delta z_{k+1} / \Delta z_k \sim C. \quad (59.4)$$

Парлетт (1964) считает, что сходимость к простому корню обычно настолько быстра, что если с какого-то момента

$$|\Delta z_{k+1} / \Delta z_k| \sim |\Delta z_k / \Delta z_{k-1}| > 0,8, \quad (59.5)$$

то можно заключить, что корень, к которому сходятся итерации, имеет по крайней мере двойную кратность, и (28.13) нужно использовать при $r = 2$. Парлетт находит, что благоразумно пользоваться этим критерием после трех итераций. Процесс может быть продолжен для попытки выявления корней высшей кратности.

Если применяется метод Ньютона с $\|A\|_\infty$ в качестве исходного приближения к каждому корню с методом удаления, описанным в § 55, то собственные значения получаются в монотонно убывающем порядке. Наконец, если мы применим последовательную линейную интерполяцию с исходными приближениями, например $(1,1) \|A\|_\infty$ и $\|A\|_\infty$, мы снова получим собственные значения в убывающем порядке. Сходимость в целом последних двух методов не настолько удовлетворительна, как сходимость метода Лагерра, и здесь следует пытаться определить лучше исходные значения. Так как собственные значения определяются в убывающем порядке, последний вычисленный корень λ_m есть верхняя граница для оставшихся корней, однако метод удаления не позволяет ее использовать. Вместо этого можно использовать $\frac{1}{2}(\lambda_m + \lambda_{m-1})$ для всех шагов после второго для метода Ньютона и $\frac{1}{3}(\lambda_{m-1} + 2\lambda_m)$ и $\frac{1}{3}(2\lambda_{m-1} + \lambda_m)$ для метода последовательной интерполяции. Если λ_{m-1} равно λ_m , то вместо него нужно взять первое из $\lambda_{m-2}, \dots, \lambda_1$ большее, чем λ_{m-1} , или, если такого нет, взять $\|A\|_\infty$. Заметим, что вместо $\|A\|_\infty$ можно использовать любую известную верхнюю границу для собственных значений. Если A была получена из исходной матрицы A_0 преобразованием подобия, вполне может случиться, что $\|A\|_\infty \geq \|A_0\|_\infty$, и в таком случае лучше использовать $\|A_0\|_\infty$. В каждом из последних двух методов не только корни вычисляются в монотонно убывающем порядке, но и сходимость к каждому корню монотонна.

Если начальное значение значительно больше любого собственного значения, то и метод Ньютона, и последовательная линейная интерполяция дают очень медленную сходимость. Использование следующего (неопубликованного) результата, открытого независимо Каганом и Маели, может привести к значительному уменьшению числа итераций. Предположим сначала, что

$$x > \lambda_1 > \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_n \quad (59.6)$$

и что мы дважды применили поправку Ньютона, так что следующее приближенное значение ξ определяется как

$$\xi = x - 2f(x)/f'(x). \quad (59.7)$$

Обозначим корни $f'(x)$ через μ_i ($i = 1, \dots, n-1$), так что

$$x > \lambda_1 > \mu_1 > \lambda_2 \geq \mu_2 \geq \lambda_3 \geq \dots \geq \lambda_{n-1} \geq \mu_{n-1} \geq \lambda_n. \quad (59.8)$$

Покажем, что $\xi > \mu_1$. Так как $f'(x) = \prod (x - \mu_i)$, ясно, что при $x \geq \mu_1$ все производные f неотрицательны. Следовательно,

$$f(x) = f(\mu_1) + \sum_2^n b_r (x - \mu_1)^r \quad (b_r \geq 0), \quad (59.9)$$

и потому

$$\begin{aligned} \xi - \mu_1 &= (x - \mu_1) - 2f(x)/f'(x) = \\ &= (x - \mu_1) - 2[f(\mu_1) + \sum_2^n b_r (x - \mu_1)^r] / \sum_2^n r b_r (x - \mu_1)^{r-1} = \\ &= [\sum_2^n (r-2) b_r (x - \mu_1)^r - 2f(\mu_1)] / \sum_2^n r b_r (x - \mu_1)^{r-1}. \end{aligned} \quad (59.10)$$

Так как $f(\mu_1) < 0$, $b_r \geq 0$, числитель и знаменатель состоят лишь из неотрицательных членов, откуда следует $\xi > \mu_1$. Заметим, что мы требовали только, чтобы λ_1 и μ_1 были вещественными и чтобы все производные f были неотрицательными при $x \geq \mu_1$. Поэтому доказательство приложимо, например, для $f(x) = x^{20} - 1$.

Аналогично при последовательной линейной интерполяции можно применить удвоенный сдвиг, полученный исходя из значений x_1 и x_2 , больших λ_1 , причем не получится значение, меньшее μ_1 . Это следует из того, что если $x_1 > x_2$, поправка, соответствующая линейной интерполяции, меньше поправки метода Ньютона, примененного в x_2 .

Следовательно, мы спокойно можем применять двойной сдвиг, и если собственное значение простое, мы не продвинемся за μ_1 . Итерации будут монотонно уменьшаться до тех пор, пока не получится значение, меньшее λ_1 . Следующий сдвиг будет положительным, и в это время прекращается использование двойного сдвига. Заметим, что доказательство остается справедливым для собственных значений любой кратности. Если, в частности, λ_1 — кратное собственное значение, метод двойного сдвига никогда не даст величины, меньшей λ_1 , и сходимость будет монотонной все время, до тех пор, пока ошибки округления не станут преобладать в вычислениях. Каган значительно расширил эти идеи и получил дальнейшие результаты в связи с методом Мюллера.

Если A — симметричная или квазисимметричная трехдиагональная матрица, то можно получить информацию о распределении нулей из свойства последовательности Штурма и деления отрезка (гл. 5, § 39). Поэтому можно комбинировать методы этой главы с методом деления. Особенно изысканная комбинация метода деления и последовательной линейной интерполяции была предложена Деккером (1962). Этот метод может использоваться для локализации вещественного корня любой вещественной функции после того, как определен интервал, содержащий этот корень.

Комплексные корни

60. Для комплексных собственных значений вещественной или комплексной матрицы уже нет удовлетворительных методов, обеспечивающих сходимость. Вообще говоря, кажется, что метод Лагерра имеет очень хорошие общие свойства сходимости, но Парлетт (1964) показал, что существуют простые полиномы, для которых некоторые начальные

значения приводят к повторяющимся циклам итераций. Он рассматривал функцию $z(z^2 + b^2)$ с $b > 0$, для которой

$$z_{k+1} = z_k - 3z_k(z_k^2 + b^2)/[3z_k^2 + b^2 \pm 2b(b^2 - 3z_k^2)^{1/2}]. \quad (60.1)$$

Если $z_1 = 3^{-1/2}b$, то $z_2 = -z_1$ и $z_3 = z_1$. Циклы возможны и во всех других методах, и, в частности, только методы Мюллера и Лагерра могут давать комплексные итерации, начиная с вещественных значений для вещественной функции. Дополнительная трудность заключается в том, что циклы могут вызваться плохой обусловленностью, а не фундаментальным срывом используемого метода. В таких случаях использование достаточно высокой точности вычислений ведет к прекращению циклов.

Трудно отрицать, что эти замечания являются серьезной критикой методов этой главы. Можно разочароваться в использовании методов, успех которых в большой степени кажется случайным, но надо помнить, что до сих пор нет других методов для общей проблемы собственных значений, которые *гарантировали бы* получение результата *в приемлемое время*. Наш опыт показал, что программы, основанные на методах этой главы, были очень эффективны и принадлежат к числу наиболее точных.

Рекомендации

61. Среди методов, использовавшихся мной для вещественных матриц, наиболее эффективный был основан на следующей комбинации методов Лагерра и Берстоу. Для каждого собственного значения сначала применялся метод Лагерра, но если сходимость не получалась за предписанное число k итераций, включался метод Берстоу. На практике k равнялось 32, хотя наиболее часто сходимость обеспечивалась 4—10 итерациями. Использовались методы удаления §§ 55, 56, а не процесс исчерпывания.

Если исходная матрица общего вида, я предпочитал привести ее к форме Хессенберга, и затем использовать эту форму для вычисления значений функции и ее производных. Вычисления с обычной точностью почти неизменно давали верными те знаки, которые не меняются при относительных возмущениях порядка 2^{-i} элементов исходной матрицы. Для общих целей этот метод оказался лучше всех других, использованных мной, за исключением метода Френсиса, описанного в следующей главе.

Я также экспериментировал с дальнейшим приведением к трехдиагональному виду, но не рекомендую, чтобы в этом случае и приведение к трехдиагональному виду, и вычисление значений функции и ее производных производились с двойной точностью для предохранения от возможного ухудшения обусловленности собственных значений. Для матриц порядка более 16 это быстрее, чем использование формы Хессенберга с обычной точностью, и за исключением редких случаев, когда ухудшение обусловленности было исключительно велико, трехдиагональные матрицы давали более точный результат. Мое предпочтение формы Хессенберга поэтому можно рассматривать как проявление излишней осторожности.

Хотя мы видели, что использование формы Фробениуса опасно, так как она может быть катастрофически плохо обусловлена по сравнению с исходной матрицей, программы, основанные на таком приведении, удивительно хороши для матриц, возникающих в задачах о затухании электрических или механических колебаний. Для соответствующих характеристических полиномов обычна хорошая обусловленность. Если это верно, методы, основанные на явных характеристических полиномах, и экономичны и точны.

Комплексные матрицы

62. Нами было рассмотрено очень мало матриц общего вида с комплексными элементами. Почти для всех матриц, рассмотренных в Национальной физической лаборатории, использовался метод Мюллера. Если исходная матрица была общего вида, снова выполнялось приведение к форме Хессенберга с использованием обычной точности. Однако в нескольких примерах исходная матрица была в трехдиагональном виде, и для них использовалось вычисление с обычной точностью, основанное на методе § 8. Для указания достигнутой точности можем упомянуть, что для группы из восьми комплексных матриц порядка 105 (наивысший порядок комплексной матрицы, с которым мы когда-либо имели дело) последующий анализ показал, что максимальная ошибка в любом собственном значении была меньше $2^{-43} \|A\|_\infty$, причем эта точность была получена при использовании мантиссы с 46 двоичными знаками. Среднее число итераций на собственное значение было равно 10.

Во всех итерационных программах для определения собственных значений верхняя граница для собственных значений вычислялась до начала итераций. Если исходная матрица имеет форму, отличную от той, которая используется для вычисления значений функции, то применяется наименьшая из $\|\cdot\|_\infty$ исходной и приведенной матриц. Эта норма дает верхнюю границу для модулей всех собственных значений; если итерация превосходит эту норму, она заменяется некоторым подходящим значением. Способ выбора этого подходящего значения не обоснован, но если не применять каких-либо приемов, может потребоваться чрезмерное количество итераций, если принятое значение лежит далеко от области, содержащей собственные значения. Например, такая ситуация возникнет, если взять $z = 1/2$ для полинома $z^{20} - 1$ в методе Ньютона. Следующее приближение равно примерно $2^{19}/20$, и возвращение к соответствующей области очень медленное.

Матрицы, содержащие независимый параметр

63. Как мы уже замечали, итерационные методы дают наилучший результат, когда приближения к собственным значениям выбираются по смыслу задачи. Существует одна важная ситуация, когда такие приближения возможно найти. Это соответствует случаю, когда элементы матрицы являются функциями независимого параметра v , и требуется определить зависимость нескольких или всех собственных значений от этого параметра.

Такая проблема возникает в связи с обобщенной проблемой собственных значений

$$\det[B_0(v) + \lambda B_1(v) + \dots + \lambda^r B_r(v)] = 0. \quad (63.1)$$

Важным примером является проблема флаттера в аэродинамике. Для этой проблемы $r = 2$, а v есть скорость самолета. Скорость флаттера это то значение v , для которого собственное значение имеет нулевую вещественную часть.

Проблема собственных значений полностью решается для начального значения скорости v_0 , а затем прослеживается зависимость некоторых или всех собственных значений для серии значений $v_0, v_1, \dots$. Если разность последовательных значений v_i достаточно мала, каждое собственное

значение при v_i является хорошим приближением для соответствующего собственного значения при v_{i+1} . Оказалось, что программа на АСЕ, основанная на методе Мюллера, в которой значения функции вычислялись прямо из $B_0(v) + \lambda B_1(v) + \dots + \lambda^r B_r(v)$ при помощи метода Гаусса с перестановками, достаточно хороша. Можно получать даже лучшие начальные приближения, дифференцируя $\lambda_s(v_i)$, но так как в требуемом интервале v сходимость обычно хорошая, это не дает должного эффекта. Без сомнения, применение линейной интерполяции было бы столь же хорошо для этой цели, но мы не имеем опыта в этом. На АСЕ проходила весьма успешно работа с обобщенной проблемой, где B_i были комплексные матрицы шестого порядка, а $r = 5$, так что при каждом v вычислялось 30 собственных значений.

Дополнительные замечания

Интерес к методу Лагерра был стимулирован Маели (1954). Вероятно, наиболее строгий анализ его эффективности на практике проведен Шарлеттом (1964), и читателю предлагается внимательно ознакомиться с его работой. Практическое использование итерационных методов подняло интерес к деталям, жизненно важным для успеха автоматических процедур, таким, как определение достижения предельной точности и распознавание кратных корней и плохой обусловленности.

В процессе напечатания этой книги Трауб (1964) опубликовал замечательное исследование об итерационных методах, которое включает улучшенную формулировку метода Мюллера, основанную на интерполяционных полиномах Ньютона, а не Лагранжа. Для квадратичного полинома в обычных обозначениях разделенных разностей имеем

$$\begin{aligned} f(w) &= f(w_i) + (w - w_i) f[w_i, w_{i-1}] + (w - w_i)(w - w_{i-1}) f[w_i, w_{i-1}, w_{i-2}] = \\ &= f(w_i) + (w - w_i) \{f[w_i, w_{i-1}] + (w_i - w_{i-1}) f[w_i, w_{i-1}, w_{i-2}]\} + \\ &\quad + (w - w_i)^2 f[w_i, w_{i-1}, w_{i-2}] = f(w_i) + (w - w_i) p + (w - w_i)^2 q, \end{aligned}$$

что дает

$$w = w_i - 2f(w_i) / \{p \pm [p^2 - 4f(w_i)q]^{1/2}\}.$$

Так как

$$f[w_i, w_{i-1}, w_{i-2}] = \{f[w_i, w_{i-1}] - f[w_{i-1}, w_{i-2}]\} / (w_i - w_{i-2}),$$

нам нужно на каждой стадии вычислять только $f[w_i, w_{i-1}]$.

LR*- и *QR*-алгоритмы*Введение**

1. В гл. 1, § 42 мы показали, что любую матрицу преобразованием подобия можно привести к треугольному виду, и далее в § 47 показали, что матрица преобразования может быть выбрана унитарной. Очевидно, что любой алгоритм, дающий такое приведение, должен быть по существу итерационным, так как он приводит к определению нулей полинома.

Как показано в гл. 2, § 48, унитарное преобразование, приводящее нормальную матрицу к *треугольному* виду, на самом деле должно приводить ее к *диагональному* виду. Метод Якоби (гл. 5, § 3) дает такое приведение в случае вещественных симметричных матриц. Основная стратегия этого метода заключается в том, что при помощи последовательных элементарных ортогональных преобразований уменьшается норма матрицы из наддиагональных элементов. При этом в силу симметрии одновременно уменьшается норма матрицы из поддиагональных элементов.

Метод Якоби может быть весьма просто распространен на все нормальные матрицы (Голдстейн и Хорвиц, 1959), и естественно пытаться распространить метод, основанный на подобной стратегии, на приведение произвольных матриц к треугольному виду. Но, несмотря на большое количество попыток, до сих пор не появилось алгоритмов, сравнимых на практике с описанными в главах 6 и 7.

Методы исчерпывания, описанные в гл. 7, §§ 49—54, эффективно осуществляют приведение матрицы Хессенберга к треугольному виду, но они численно неустойчивы. В главе 9 будут описаны устойчивые методы исчерпывания, которые обеспечат подобное приведение произвольной матрицы к треугольному виду.

В этой главе мы займемся двумя алгоритмами, известными под названиями *LR*- и *QR*-алгоритмов. Первый из них, развитый Рутисхаузером, приводит произвольную матрицу к треугольному виду при помощи неунитарных преобразований. По моему мнению, создание этого метода является наиболее значительным вкладом в решение проблемы собственных значений, сделанным после появления автоматических вычислительных машин. Алгоритм *QR*, развитый позже Френсисом *), тесно связан с *LR*-алгоритмом, но основан на использовании унитарных преобразований. Во многих отношениях он является наиболее эффективным из известных методов решения общей алгебраической проблемы собственных значений.

*) *QR*-алгоритм одновременно был предложен В. Н. Кублаповской (1961). Ею в 1962 г. впервые была доказана сходимость метода для произвольной матрицы. (Прим. перев.)

Вещественные матрицы с комплексными собственными значениями

2. Неудобством приведения к матрице треугольного вида является то, что если вещественная матрица имеет несколько комплексно сопряженных собственных значений, то такое приведение переводит нас в поле комплексных чисел. Мы введем простую модификацию треугольной формы, которая позволит оставаться в поле вещественных чисел. Покажем, что если вещественная матрица A имеет собственные значения $\lambda_1 \pm i\mu_1, \lambda_2 \pm i\mu_2, \dots, \lambda_s \pm i\mu_s, \lambda_{2s+1}, \dots, \lambda_n$, где λ_i и μ_i — вещественные, то существует такая вещественная матрица H , что матрица $B = HAH^{-1}$ имеет вид

$$B = \begin{bmatrix} \boxed{X_1} & & & & & \\ & \boxed{X_2} & & & & \\ & & \ddots & & & \\ & & & \boxed{X_s} & & \\ & & & & \lambda_{2s+1} & \\ & & & & & \ddots \\ & & & & & & \lambda_n \end{bmatrix}, \quad (2.1)$$

где X_r это матрицы второго порядка, стоящие на диагонали B и имеющие собственные значения $\lambda_r \pm i\mu_r$, а остальные поддиагональные элементы B равны нулю. Матрица B , следовательно, отличается от треугольной тем, что она имеет s вещественных поддиагональных элементов, по одному связанному с каждой X_r . Покажем далее, что H может быть выбрана ортогональной.

Доказательство аналогично доказательству, приведенному в гл. 1, §§ 42, 47 для унитарного и неунитарного приведения к треугольному виду. Единственное отличие возникает из-за присутствия пар комплексно сопряженных собственных значений. В силу доказательства из главы 1 достаточно показать, что существует такая матрица H_1 , что

$$H_1 A H_1^{-1} = \begin{bmatrix} \boxed{X_1} & \boxed{P_1} \\ 0 & \boxed{A_2} \end{bmatrix}, \quad (2.2)$$

так как дальше результат будет следовать по индукции. Мы можем показать это следующим образом.

Пусть $x_1 \pm iy_1$ — собственные векторы, соответствующие $\lambda_1 \pm i\mu_1$. Очевидно, что x_1 и y_1 должны быть линейно независимыми, так как иначе соответствующее собственное значение было бы вещественным, и мы имеем

$$A [x_1 | y_1] = [x_1 | y_1] \begin{bmatrix} \lambda_1 & \mu_1 \\ -\mu_1 & \lambda_1 \end{bmatrix} = [x_1 | y_1] \Lambda_1. \quad (2.3)$$

Предположим, что можно найти такую неособенную H_1 , что

$$H_1 [x_1 | y_1] = \begin{bmatrix} V_1 \\ 0 \end{bmatrix}, \quad (2.4)$$

где V_1 — неособенная матрица второго порядка. Тогда из (2.3) имеем

$$H_1 A H_1^{-1} H_1 [x_1 \mid y_1] = H_1 [x_1 \mid y_1] \Lambda_1, \quad (2.5)$$

что дает

$$H_1 A H_1^{-1} \begin{bmatrix} V_1 \\ 0 \end{bmatrix} = \begin{bmatrix} V_1 \\ 0 \end{bmatrix} \Lambda_1. \quad (2.6)$$

Следовательно, если положить

$$H_1 A H_1^{-1} = \left[\begin{array}{c|c} X_1 & P_1 \\ \hline Q_1 & A_2 \end{array} \right], \quad (2.7)$$

то уравнение (2.6) даст

$$X_1 V_1 = V_1 \Lambda_1, \quad Q_1 V_1 = 0, \quad (2.8)$$

откуда видно, что $Q_1 = 0$ и что X_1 подобна Λ_1 и поэтому имеет собственные значения $\lambda_i \pm i\mu_i$. Любая H_1 , для которой выполнено равенство (2.4), удовлетворяет, следовательно, нашим требованиям.

Построить такую H_1 можно либо в виде произведения вещественных устойчивых элементарных неунитарных матриц, либо в виде произведения вещественных плоских вращений или вещественных матриц отражения. Это становится очевидным, если взять x_1 и y_1 в качестве первых двух столбцов матрицы n -го порядка и рассмотреть ее приведение к треугольному виду по методу Гаусса с выбором главного элемента по столбцу или по методам Гивенса или Хаусхолдера. В каждом из этих случаев V_1 имеет вид

$$\begin{bmatrix} \times & \times \\ 0 & \times \end{bmatrix}, \quad (2.9)$$

и ее неособенность следует из линейной независимости x_1 и y_1 .

LR-алгоритм

3. Алгоритм Рунтисхаузера основан на разложении матрицы в произведение треугольных (гл. 4, § 36). По Рунтисхаузеру (1958)

$$A = LR, \quad (3.1)$$

где L — единичная нижняя треугольная, а R — верхняя треугольная. В этой главе используется обозначение $\mathcal{R}$ вместо U для того, чтобы облегчить сравнение с работой Рунтисхаузера.

Предположим, что мы составим преобразование подобия $L^{-1}AL$ матрицы A . Тогда

$$L^{-1}AL = L^{-1}(LR)L = RL. \quad (3.2)$$

Следовательно, если разложить A в произведение треугольных, а затем перемножить сомножители в обратном порядке, то получится матрица, подобная A . В LR-алгоритме этот процесс бесконечно повторяется. Если исходную матрицу переобозначить через A_1 , то алгоритм определится уравнениями

$$A_{s-1} = L_{s-1}R_{s-1}, \quad R_{s-1}L_{s-1} = A_s. \quad (3.3)$$

Очевидно, что A_s подобна A_{s-1} и, следовательно, по индукции подобна A_1 . Рутисхаузер показал, что при некоторых ограничениях

$$L_s \rightarrow I \text{ и } R_s \rightarrow A_s \rightarrow \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & X \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \text{ при } s \rightarrow \infty. \quad (3.4)$$

4. Прежде чем доказывать этот результат, установим некоторые соотношения между последовательными итерациями, которые будут постоянно нужны. Из (3.3) имеем

$$A_s = L_{s-1}^{-1} A_{s-1} L_{s-1}, \quad (4.1)$$

и повторное применение дает

$$A_s = L_{s-1}^{-1} L_{s-2}^{-1} \dots L_2^{-1} L_1^{-1} A_1 L_1 L_2 \dots L_{s-1}, \quad (4.2)$$

или

$$L_1 L_2 \dots L_{s-1} A_s = A_1 L_1 L_2 \dots L_{s-1}. \quad (4.3)$$

Матрицы

$$T_s = L_1 L_2 \dots L_s \quad \text{и} \quad U_s = R_s R_{s-1} \dots R_1 \quad (4.4)$$

это *единичная нижняя треугольная* и *верхняя треугольная* соответственно.

Рассмотрим произведение $T_s U_s$. Имеем

$$\begin{aligned} T_s U_s &= L_1 L_2 \dots L_{s-1} (L_s R_s) R_{s-1} \dots R_2 R_1 = L_1 L_2 \dots L_{s-1} A_s R_{s-1} \dots R_2 R_1 = \\ &= A_1 L_1 L_2 \dots L_{s-1} R_{s-1} \dots R_2 R_1 = A_1 T_{s-1} U_{s-1}. \end{aligned} \quad (4.5)$$

Повторно применяя этот результат, получим

$$T_s U_s = A_1^s, \quad (4.6)$$

так что $T_s U_s$ дает разложение матрицы A_1^s в произведение двух треугольных.

Доказательство сходимости последовательности A_s

5. Соотношение (4.6) является фундаментальным результатом, на котором основан весь анализ. Используем его для доказательства того, что если собственные значения λ_i матрицы A удовлетворяют соотношениям

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|, \quad (5.1)$$

то, вообще говоря, результат (3.4) справедлив. Прежде чем давать формальное доказательство, рассмотрим простой пример матрицы третьего порядка. Если для простоты матрицу X правых собственных векторов записать в виде

$$X = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix} \quad (5.2)$$

и ее обратную Y в виде

$$X^{-1} = Y = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}, \quad (5.3)$$

то

$$F = A_1^s = X \begin{bmatrix} \lambda_1^s \\ \lambda_2^s \\ \lambda_3^s \end{bmatrix} X^{-1} = \quad (5.4)$$

$$= \begin{bmatrix} \lambda_1^s x_1 a_1 + \lambda_2^s y_1 a_2 + \lambda_3^s z_1 a_3 & \lambda_1^s x_1 b_1 + \lambda_2^s y_1 b_2 + \lambda_3^s z_1 b_3 & \lambda_1^s x_1 c_1 + \lambda_2^s y_1 c_2 + \lambda_3^s z_1 c_3 \\ \lambda_1^s x_2 a_1 + \lambda_2^s y_2 a_2 + \lambda_3^s z_2 a_3 & \lambda_1^s x_2 b_1 + \lambda_2^s y_2 b_2 + \lambda_3^s z_2 b_3 & \lambda_1^s x_2 c_1 + \lambda_2^s y_2 c_2 + \lambda_3^s z_2 c_3 \\ \lambda_1^s x_3 a_1 + \lambda_2^s y_3 a_2 + \lambda_3^s z_3 a_3 & \lambda_1^s x_3 b_1 + \lambda_2^s y_3 b_2 + \lambda_3^s z_3 b_3 & \lambda_1^s x_3 c_1 + \lambda_2^s y_3 c_2 + \lambda_3^s z_3 c_3 \end{bmatrix}. \quad (5.5)$$

Так как $T_s U_s$ есть разложение A_1^s в произведение треугольных, то элементы первого столбца T_s равны

$$\left. \begin{aligned} t_{11}^{(s)} &= 1, \\ t_{21}^{(s)} &= (\lambda_1^s x_2 a_1 + \lambda_2^s y_2 a_2 + \lambda_3^s z_2 a_3) / (\lambda_1^s x_1 a_1 + \lambda_2^s y_1 a_2 + \lambda_3^s z_1 a_3), \\ t_{31}^{(s)} &= (\lambda_1^s x_3 a_1 + \lambda_2^s y_3 a_2 + \lambda_3^s z_3 a_3) / (\lambda_1^s x_1 a_1 + \lambda_2^s y_1 a_2 + \lambda_3^s z_1 a_3). \end{aligned} \right\} \quad (5.6)$$

Очевидно, что если $x_1 a_1 \neq 0$, то

$$t_{21}^{(s)} = x_2/x_1 + O(\lambda_2/\lambda_1)^s \quad \text{и} \quad t_{31}^{(s)} = x_3/x_1 + O(\lambda_2/\lambda_1)^s, \quad (5.7)$$

и элементы T_s стремятся, вообще говоря, к соответствующим элементам, полученным при разложении X в произведение треугольных.

Для элементов второго столбца T_s аналогично имеем

$$\left. \begin{aligned} t_{22}^{(s)} &= 1, \\ t_{32}^{(s)} &= (f_{11}f_{32} - f_{12}f_{31}) / (f_{11}f_{22} - f_{12}f_{21}) = \\ &= \frac{(\lambda_1 \lambda_2)^s (x_1 y_3 - x_3 y_1) (a_1 b_2 - a_2 b_1) + \dots}{(\lambda_1 \lambda_2)^s (x_1 y_2 - x_2 y_1) (a_1 b_2 - a_2 b_1) + \dots} = \frac{x_1 y_3 - x_3 y_1}{x_1 y_2 - x_2 y_1} + O\left(\frac{\lambda_3}{\lambda_2}\right)^s + \dots \end{aligned} \right\} \quad (5.8)$$

при условии, что

$$(a_1 b_2 - a_2 b_1) (x_1 y_2 - x_2 y_1) \neq 0. \quad (5.9)$$

Из (5.8) мы видим, что предельное значение $t_{32}^{(s)}$ равно соответствующему элементу, полученному при разложении X в произведение треугольных. Поэтому мы установили, что если

$$X = TU \quad (5.10)$$

при $x_1 a_1 \neq 0$ и $(x_1 y_2 - x_2 y_1)(a_1 b_2 - a_2 b_1) \neq 0$, то

$$T_s \rightarrow T. \quad (5.11)$$

6. Мы показали в этом простом случае, что если ведущие главные миноры матриц X и Y не равны нулю, то матрицы T_s стремятся к единичной нижней треугольной матрице, полученной при разложении матрицы собственных векторов X в произведение треугольных. Покажем теперь, что этот результат справедлив вообще для матриц с различными собственными значениями. Если положить

$$T_s U_s = A^s = B, \quad (6.1)$$

то из явных выражений для элементов треугольных сомножителей матрицы, полученных в гл. 4, § 19, имеем

$$t_{ji}^{(s)} = \det(B_{ji}) / \det(B_{ii}). \quad (6.2)$$

Здесь B_{ii} обозначает ведущую главную подматрицу B , а B_{ji} есть подматрица, полученная из B_{ii} заменой элементов i -й строки соответствующими элементами j -й строки B . Из соотношения

$$B = A^s = X \operatorname{diag}(\lambda_i^s) Y \quad (6.3)$$

имеем

$$B_{ji} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{i-1,1} & x_{i-1,2} & \dots & x_{i-1,n} \\ x_{j1} & x_{j2} & \dots & x_{jn} \end{bmatrix} \times \begin{bmatrix} \lambda_1^s y_{11} & \lambda_1^s y_{12} & \dots & \lambda_1^s y_{1i} \\ \lambda_2^s y_{21} & \lambda_2^s y_{22} & \dots & \lambda_2^s y_{2i} \\ \dots & \dots & \dots & \dots \\ \lambda_n^s y_{n1} & \lambda_n^s y_{n2} & \dots & \lambda_n^s y_{ni} \end{bmatrix}. \quad (6.4)$$

Согласно теореме Бине — Коши (гл. 1, § 15) $\det(B_{ji})$ равен сумме произведений соответствующих i -строчных миноров двух матриц из правой части (6.4). Следовательно, можно написать

$$t_{ji}^{(s)} = \frac{\sum x_{p_1 p_2}^{(j)} \dots p_i y_{p_1 p_2}^{(i)} \dots p_i (\lambda_{p_1} \lambda_{p_2} \dots \lambda_{p_i})^s}{\sum x_{p_1 p_2}^{(i)} \dots p_i y_{p_1 p_2}^{(i)} \dots p_i (\lambda_{p_1} \lambda_{p_2} \dots \lambda_{p_i})^s}, \quad (6.5)$$

где $x_{p_1 p_2}^{(j)} \dots p_i$ это i -строчный минор, состоящий из строк 1, 2, ..., ..., $i-1$, j и столбцов $p_1, p_2, \dots, p_i$ матрицы X , а $y_{p_1 p_2}^{(i)} \dots p_i$ это i -строчный минор, состоящий из строк $p_1, p_2, \dots, p_i$ и столбцов 1, 2, ..., i матрицы Y .

Главные члены в числителе и знаменателе (6.5) определяются множителем $(\lambda_1 \lambda_2 \dots \lambda_i)^s$ при условии, что соответствующие коэффициенты не равны нулю. Этот член в знаменателе равен

$$\det(X_{ii}) \det(Y_{ii}) (\lambda_1 \lambda_2 \dots \lambda_i)^s, \quad (6.6)$$

где X_{ii} и Y_{ii} — главные ведущие подматрицы порядка i . Предполагая, что $\det(X_{ii}) \det(Y_{ii}) \neq 0$, имеем

$$t_{ji}^{(s)} \rightarrow \frac{\det(X_{ji}) \det(Y_{ii})}{\det(X_{ii}) \det(Y_{ii})} = \frac{\det(X_{ji})}{\det(X_{ii})}, \quad (6.7)$$

что показывает, что $T_s \rightarrow T$, где

$$X = TU. \quad (6.8)$$

Из (4.3) и (4.4) имеем

$$A_s = T_{s-1}^{-1} A_1 T_{s-1} \rightarrow T^{-1} A_1 T = UX^{-1}AXU^{-1} = U \operatorname{diag}(\lambda_i) U^{-1}, \quad (6.9)$$

откуда следует, что предельная матрица для A_s есть верхняя треугольная с диагональными элементами λ_i .

Используя соотношение $L_s = T_{s-1} T_s$ и уравнение (6.5), можно элементарным, но весьма утомительным способом доказать, что

$$l_{ij}^{(s)} = O(\lambda_i / \lambda_j)^s \quad (s \rightarrow \infty). \quad (6.10)$$

Отсюда, используя соотношение $A_s = L_s R_s$ и тот факт, что R_s стремится к пределу, можем вывести, что

$$a_{ij}^{(s)} = O(\lambda_i/\lambda_j)^s \quad (s \rightarrow \infty) \quad (i > j). \quad (6.11)$$

Следовательно, если некоторые собственные значения плохо разделены, то сходимость A_s к треугольной матрице может быть медленной. Другое (и более простое) доказательство дано в § 33, но метод доказательства, используемый здесь, более поучителен.

7. При получении этих результатов явно или неявно были сделаны следующие предположения:

(i) Все собственные значения различны по абсолютной величине. Заметим, что A может быть комплексной, но не может быть вещественной матрицей, имеющей комплексно сопряженные собственные значения (последний случай обсуждается в § 9).

(ii) Разложение в произведение треугольных матриц существует на каждом этапе. Легко построить матрицы, которые не являются исключительными в других отношениях, для которых это условие не удовлетворяется, например

$$A = \begin{bmatrix} 0 & 1 \\ -3 & 4 \end{bmatrix}. \quad (7.1)$$

Эта матрица имеет собственные значения 1 и 3, но не может быть разложена в произведение треугольных. Однако заметим, что, например, $(A + I)$ может быть разложена в произведение треугольных, и мы можем работать с этой модифицированной матрицей.

(iii) Ведущие главные миноры X и Y не равны нулю. Существуют очень простые матрицы, для которых это требование не выполнено. Наиболее очевидными примерами таких матриц являются матрицы вида

$$\left[\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array} \right], \quad (7.2)$$

которые сохраняют свою форму на каждом этапе. Матрицы A_1 и A_2 обрабатываются независимо, и предельная матрица имеет собственные значения A_1 в верхнем левом углу и собственные значения A_2 в нижнем правом углу независимо от их относительной величины. Полезно рассмотреть матрицу второго порядка

$$\begin{bmatrix} a & \varepsilon_1 \\ \varepsilon_2 & b \end{bmatrix}, \quad (7.3)$$

где $|b| > |a|$. Если ε_i равны нулю, то матрица не меняется при применении LR-алгоритма, и предельная матрица не будет иметь собственные значения a и b расположенными в правильном порядке. Матрица собственных векторов равна $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, и ее первый главный минор равен нулю.

Если ε_i не нули, но малы, то собственные значения очень близки к a и b . Предельная матрица теперь имеет собственные значения, расположенные в правильном порядке, но не удивительно, что сходимость может быть медленной, даже если a и b не особенно близки.

Ясно, что в этом примере две половины матрицы полностью «не связаны» при $\varepsilon_1 = \varepsilon_2 = 0$, и введение малых ε_1 и ε_2 приводит к появлению

«слабой связи». В более сложных примерах мы первоначально можем иметь полную «несвязанность», т. е. при некотором i иметь

$$\det(X_{ii})\det(Y_{ii}) = 0, \quad (7.4)$$

но ошибки округления в LR -процессе могут внести эффект слабой связи.

Положительно определенные эрмитовы матрицы

8. Если A_1 — положительно определенная эрмитова матрица, то можно снять ограничения § 7. В этом случае

$$A_1 = X \operatorname{diag}(\lambda_i) X^H, \quad (8.1)$$

где X — унитарная, а λ_i — вещественные и положительные. Следовательно, уравнение (6.5) перейдет в

$$t_{ji}^{(s)} = \frac{\sum x_{p_1 p_2 \dots p_i}^{(j)} \overline{x_{p_1 p_2 \dots p_i}^{(i)}} (\lambda_{p_1} \lambda_{p_2} \dots \lambda_{p_i})^s}{\sum |x_{p_1 p_2 \dots p_i}^{(i)}|^2 (\lambda_{p_1} \lambda_{p_2} \dots \lambda_{p_i})^s}. \quad (8.2)$$

Рассмотрим класс наборов $p_1 p_2 \dots p_i$, для которых $x_{p_1 p_2 \dots p_i}^{(i)} \neq 0$. Пусть $q_1 q_2 \dots q_i$ — такой представитель этого класса, что $\lambda_{q_1} \lambda_{q_2} \dots \lambda_{q_i}$ принимает наибольшее значение. Тогда в знаменателе преобладает член (или члены) с $(\lambda_{q_1} \lambda_{q_2} \dots \lambda_{q_i})^s$. Что касается числителя, то мы знаем из определения $q_1 q_2 \dots q_i$, что не может существовать член более высокого порядка, чем $(\lambda_{q_1} \lambda_{q_2} \dots \lambda_{q_i})^s$, и, конечно, этот член может иметь нулевой коэффициент. Следовательно, $t_{ij}^{(s)}$ стремится к пределу, который может быть нулем.

Сейчас мы не хотим делать никаких предположений относительно главных миноров X , и поэтому не можем утверждать, что матрица, предельная для T_s , есть матрица, полученная при разложении X в произведение треугольных. Соответственно не будем действовать так, как в (6.9), а напомним

$$T_s \rightarrow T_\infty, \quad (8.3)$$

так что

$$L_s = T_{s-1}^{-1} T_s \rightarrow T_\infty^{-1} T_\infty = I. \quad (8.4)$$

Далее имеем

$$A_s = T_{s-1}^{-1} A_1 T_{s-1} \rightarrow T_\infty^{-1} A_1 T_\infty, \quad (8.5)$$

так что матрицы A_s стремятся к пределу, который мы назовем A_∞ . Теперь

$$R_s = L_s^{-1} A_s \rightarrow I T_\infty^{-1} A_1 T_\infty, \quad (8.6)$$

и, следовательно, матрицы R_s стремятся к тому же пределу, что и A_s . Этот предел необходимо будет треугольной матрицей, так как при всех s матрицы R_s треугольные. Ее диагональными элементами должны быть собственные значения A_1 , расположенные в некотором порядке, так как по (8.5) она подобна A_1 .

Заметим, что на это доказательство не влияют ни присутствие кратных собственных значений, ни обращение в нуль некоторых главных ведущих миноров X . Собственные значения λ_i не обязательно располагаются на диагонали предельной матрицы в убывающем порядке. Снова

существует опасность, что в случае, когда точные вычисления вели бы к предельной матрице с неправильно упорядоченными собственными значениями, влияние ошибок округления может привести к *очень медленной* сходимости к треугольной матрице, собственные значения которой упорядочены правильно.

Комплексно сопряженные собственные значения

9. Если A_1 — вещественная, но имеет одну или более пар комплексно сопряженных собственных значений, то очевидно, что матрицы A_s не могут стремиться к верхней треугольной матрице, имеющей эти собственные значения на диагонали, так как все A_s — вещественные. Рассмотрим сначала случай, когда имеется одна пара комплексно сопряженных собственных значений λ_{m-1} и $\bar{\lambda}_{m-1}$, а для остальных собственных значений $|\lambda_i| > |\lambda_j|$ при $i < j$. Будем предполагать, что ни один из ведущих миноров X и Y не равен нулю. Из (6.5) следует, что

$$t_{ji}^{(s)} \rightarrow \det X_{ji} / \det X_{ii} \quad (i \neq m-1) \quad (9.1)$$

точно так же, как и в предыдущем случае. Однако для $i = m-1$ в числителе и знаменателе имеются члены как с $(\lambda_1 \lambda_2 \dots \lambda_{m-2} \lambda_{m-1})^s$, так и с $(\lambda_1 \lambda_2 \dots \lambda_{m-2} \bar{\lambda}_{m-1})^s$. Для изучения асимптотического поведения таких элементов удобно ввести упрощенные обозначения для миноров порядка $(m-1)$ матриц X и Y . Положим

$$p_j = \det \begin{bmatrix} x_{11} & \dots & x_{1, m-1} \\ x_{21} & \dots & x_{2, m-1} \\ \dots & \dots & \dots \\ x_{m-2, 1} & \dots & x_{m-2, m-1} \\ x_{j1} & \dots & x_{j, m-1} \end{bmatrix}, \quad q_{m-1} = \det \begin{bmatrix} y_{11} & \dots & y_{1, m-1} \\ y_{21} & \dots & y_{2, m-1} \\ \dots & \dots & \dots \\ y_{m-1, 1} & \dots & y_{m-1, m-1} \end{bmatrix}. \quad (9.2)$$

Вспомня, что столбцы $(m-1)$ и m матрицы X и строки $(m-1)$ и m матрицы Y являются комплексно сопряженными, имеем

$$\bar{p}_j = \det \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1, m-2} & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2, m-2} & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m-2, 1} & x_{m-2, 2} & \dots & x_{m-2, m-2} & x_{m-2, m} \\ x_{j1} & x_{j2} & \dots & x_{j, m-2} & x_{jm} \end{bmatrix},$$

$$\bar{q}_{m-1} = \det \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1, m-1} \\ y_{21} & y_{22} & \dots & y_{2, m-1} \\ \dots & \dots & \dots & \dots \\ y_{m-2, 1} & y_{m-2, 2} & \dots & y_{m-2, m-1} \\ y_{m1} & y_{m2} & \dots & y_{m, m-1} \end{bmatrix}. \quad (9.3)$$

Уравнение (6.5) теперь дает

$$t_{j, m-1}^{(s)} \sim \frac{p_j q_{m-1} (\lambda_1 \lambda_2 \dots \lambda_{m-1})^s + \bar{p}_j \bar{q}_{m-1} (\lambda_1 \lambda_2 \dots \lambda_{m-2} \bar{\lambda}_{m-1})^s}{p_{m-1} q_{m-1} (\lambda_1 \lambda_2 \dots \lambda_{m-1})^s + \bar{p}_{m-1} \bar{q}_{m-1} (\lambda_1 \lambda_2 \dots \lambda_{m-2} \bar{\lambda}_{m-1})^s}, \quad (9.4)$$

причем видно, что последовательность $t_{j, m-1}^{(s)}$ не сходится при $s \rightarrow \infty$. Мы можем записать (9.4) в виде

$$t_{j, m-1}^{(s)} \sim \frac{a_s p_j + \bar{a}_s \bar{p}_j}{a_s p_{m-1} + \bar{a}_s \bar{p}_{m-1}}, \quad (9.5)$$

тогда

$$\begin{aligned} t_{j, m-1}^{(s+1)} - t_{j, m-1}^{(s)} &\sim \frac{(p_{m-1} \bar{p}_j - \bar{p}_{m-1} p_j) (\bar{a}_{s+1} a_s - a_{s+1} \bar{a}_s)}{(a_{s+1} p_{m-1} + \bar{a}_{s+1} \bar{p}_{m-1}) (a_s p_{m-1} + \bar{a}_s \bar{p}_{m-1})} = \\ &= P_s (p_{m-1} \bar{p}_j - \bar{p}_{m-1} p_j). \end{aligned} \quad (9.6)$$

Поэтому разность $(m-1)$ -х столбцов T_{s+1} и T_s есть вектор, параллельный фиксированному вектору с компонентами $p_{m-1} \bar{p}_j - \bar{p}_{m-1} p_j$. Из определения $p_j, \bar{p}_j, p_{m-1}$ и $\bar{p}_{m-1}$ и теоремы Сильвестра (см., например, Гантмахер (1967)) имеем:

$$p_{m-1} \bar{p}_j - \bar{p}_{m-1} p_j = (x_{12} \dots x_{m-2}) (x_{12}^{(j)} \dots x_{m-2}^{(j)}). \quad (9.7)$$

Первый сомножитель не зависит от j , и так как

$$\lim t_{jm}^{(s)} = x_{12}^{(j)} \dots x_{m-2}^{(j)} / x_{12}^{(m)} \dots x_{m-2}^{(m)}, \quad (9.8)$$

можно написать

$$t_{j, m-1}^{(s+1)} - t_{j, m-1}^{(s)} \sim Q_s t_{jm}^{(s)}. \quad (9.9)$$

10. Итак, доказано, что все столбцы T_s , кроме $(m-1)$ -го, стремятся к пределу, а разность $(m-1)$ -х столбцов T_{s+1} и T_s отличается лишь множителем от m -го столбца T_s . Следовательно, так как $T_{s+1} = T_s L_{s+1}$, то

$$L_{s+1} \sim \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & \boxed{Q_s \quad 1} & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix}, \quad (10.1)$$

где элемент Q_s находится в позиции $(m, m-1)$. Из соотношения $T_{s+1}^{-1} = L_{s+1}^{-1} T_s^{-1}$ следует, что все строки T_s^{-1} стремятся к пределу, кроме m -й, а m -я строка T_{s+1}^{-1} отличается от m -й строки T_s^{-1} на $(m-1)$ -ю строку T_s^{-1} , умноженную на $-Q_s$. Следовательно, последовательные $A_{s+1} = T_{s+1}^{-1} A_1 T_s$ стремятся к пределу, за исключением их m -й строки и $(m-1)$ -го столбца. Далее, L_{s+1} получается при разложении A_{s+1} в произведение треугольных, и из предельной формы L_{s+1} видно, что либо все поддиагональные элементы A_{s+1} , кроме $(m, m-1)$ -го, стремятся к нулю, либо некоторые из диагональных элементов A_{s+1} стремятся к бесконечности. Из (9.4) $t_{j, m-1}^{(s)}$ могут быть выражены в виде

$$t_{j, m-1}^{(s)} \sim R \cos(s\theta + \alpha_j + \beta_{m-1}) / \cos(s\theta + \alpha_{m-1} + \beta_{m-1}). \quad (10.2)$$

Выражение в правой части может принимать произвольно большие значения, если только θ не представимо в виде $d\pi/e$, где d и e — целые числа. Однако оно также принимает значения порядка R для бесконечного числа значений s . Следовательно, мы можем исключить возможность стремления элементов A_{s+1} к бесконечности.

11. Обобщая очевидным образом наши рассуждения, можно получить следующие результаты.

Пусть A — вещественная матрица, и пусть собственные значения пронумерованы так, что $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$, причем типичное вещественное собственное значение обозначается через λ_r , а типичная комплексно сопряженная пара — через λ_{c-1} и λ_c , так что $\lambda_c = \bar{\lambda}_{c-1}$. Если:

- (i) $|\lambda_i| > |\lambda_{i+1}|$, за исключением комплексно сопряженных пар;
- (ii) $\det(X_{ii}) \det(Y_{ii}) \neq 0$, ($i = 1, \dots, n$);
- (iii) на каждом шаге A_r имеет разложение в произведение треугольных, то:

- (i) $a_{ij}^{(s)} \rightarrow 0$ ($i > j$), за исключением $a_{c,c-1}^{(s)}$;
- (ii) $a_{rr}^{(s)} \rightarrow \lambda_r$;
- (iii) все элементы $a_{ij}^{(s)}$ ($i > i$) стремятся к пределу, за исключением элементов строк c и столбцов ($c - 1$);
- (iv) матрицы второго порядка

$$\begin{bmatrix} a_{c-1, c-1}^{(s)} & a_{c-1, c}^{(s)} \\ a_{c, c-1}^{(s)} & a_{c, c}^{(s)} \end{bmatrix}$$

не сходятся, но их собственные значения сходятся к соответствующим λ_{c-1} и λ_c ;

- (v) элементы $t_{ij}^{(s)}$ стремятся к пределам (6.7), за исключением элементов столбцов ($c - 1$).

Заметим, что LR-преобразование эффективно осуществляет приведение вещественной матрицы к виду, описанному в § 2.

12. Мы можем теперь попытаться оценить значение LR-алгоритма как практического приема. На первый взгляд он не слишком много обещающий по следующим причинам:

(i) Существуют матрицы, не имеющие разложения в произведение треугольных, несмотря на то, что для них проблема собственных значений хорошо обусловлена. Для таких матриц LR-алгоритм неприменим без некоторых модификаций. Кроме того, существует значительно более широкий класс матриц, для которых разложение в произведение треугольных численно неустойчиво. Численная неустойчивость может возникать на каждом шаге итерационного процесса и вести к значительной потере точности вычисленных собственных значений.

(ii) Объем вычислений очень велик. Каждый шаг итерации требует $\frac{2}{3}n^3$ умножений, половина из них при разложении в произведение треугольных, а остальные при последующем перемножении.

(iii) Сходимость поддиагональных элементов к нулю зависит от отношений $(\lambda_{r+1}/\lambda_r)$, и может быть очень медленной, если собственные значения плохо отделены.

Поэтому если LR-алгоритм должен конкурировать с лучшими из методов, описанных в главах 6 и 7, то его нужно так модифицировать, чтобы он выдержал эту критику.

Введение перестановок

13. Численная устойчивость разложения в произведение двух треугольных матриц может быть улучшена введением там, где это необходимо, перестановок. Этот прием уже использовался при подобном преобразовании к форме Хессенберга. Рассмотрим аналогичную модификацию LR -алгоритма.

Если A — произвольная матрица, то, как мы видели в гл. 4, § 21, существуют такие матрицы $I_{r,r'}$ и N_r , что

$$N_{n-1}I_{n-1,(n-1)'} \dots N_2I_{2,2'}N_1I_{1,1'}A = R, \quad (13.1)$$

где все элементы $|N_r|$ не более единицы. Преобразование подобия A заканчивается умножением R справа на

$$I_{1,1'}N_1^{-1}I_{2,2'}N_2^{-1} \dots I_{n-1,(n-1)'}N_{n-1}^{-1},$$

так что

$$N_{n-1}I_{n-1,(n-1)'} \dots N_2I_{2,2'}N_1I_{1,1'}AI_{1,1'}N_1^{-1}I_{2,2'}N_2^{-1} \dots I_{n-1,(n-1)'}N_{n-1}^{-1} = \\ = RI_{1,1'}N_1^{-1}I_{2,2'}N_2^{-1} \dots I_{n-1,(n-1)'}N_{n-1}^{-1}. \quad (13.2)$$

В частном случае, когда перестановки не требуются и $I_{rr'} = I$ ($r = 1, \dots, n-1$), произведение матриц, предшествующее A в (13.1), это матрица L^{-1} , для которой $L^{-1}A = R$, так что $A = LR$. В этом случае матрица в правой части (13.2) есть RL . Соответственно предлагаем следующую модификацию LR -алгоритма.

На каждом этапе матрица A_s приводится к верхней треугольной матрице R_s методом Гаусса с перестановками. Затем R_s умножается справа на сомножители, обратные к используемым при приведении. Это дает нам матрицу A_{s+1} . Объем вычислений такой же, как и при обычном LR -процессе.

Для того чтобы избежать утомительных выкладок, проанализируем связь между A_1 и A_2 более подробно только в случае $n = 4$.

Имеем

$$N_3I_{3,3'}N_2I_{2,2'}N_1I_{1,1'}A_1 = R_1, \\ A_2 = \{N_3I_{3,3'}N_2I_{2,2'}N_1I_{1,1'}\} A_1 \{I_{1,1'}N_1^{-1}I_{2,2'}N_2^{-1}I_{3,3'}N_3^{-1}\} = \\ = \{N_3I_{3,3'}N_2(I_{3,3'}I_{3,3'})I_{2,2'}N_1(I_{2,2'}I_{3,3'}I_{3,3'}I_{2,2'})I_{1,1'}\} \times \\ \times A_1 \{I_{1,1'}(I_{2,2'}I_{3,3'}I_{3,3'}I_{2,2'})N_1^{-1}I_{2,2'}(I_{3,3'}I_{3,3'})N_2^{-1}I_{3,3'}N_3^{-1}\}, \quad (13.3)$$

где все члены в круглых скобках равны I . Обозначим

$$I_{3,3'}I_{2,2'}N_1I_{2,2'}I_{3,3'} = \tilde{N}_1, \quad I_{3,3'}N_2I_{3,3'} = \tilde{N}_2, \quad (13.4)$$

причем $\tilde{N}_1$ и $\tilde{N}_2$ имеют ту же форму, что и N_1 и N_2 , но их поддиагональные элементы расположены в другом порядке. После перегруппировки сомножителей (13.3) дает

$$A_2 = \tilde{L}_1^{-1}\tilde{A}_1\tilde{L}_1, \quad \tilde{L}_1 = \tilde{N}_1^{-1}\tilde{N}_2^{-1}\tilde{N}_3^{-1}, \quad \tilde{A}_1 = I_{3,3'}I_{2,2'}I_{1,1'}A_1I_{1,1'}I_{2,2'}I_{3,3'}. \quad (13.5)$$

Очевидно, $\tilde{L}_1$ — единичная нижняя треугольная матрица, а $\tilde{A}_1$ это A_1 , у которой переставлены строки и столбцы. Если положить

$$I_{3,3'} I_{2,2'} I_{1,1'} = P_1, \quad (13.6)$$

где P_1 — матрица перестановок, то

$$\tilde{A}_1 = P_1 A_1 P_1^T, \quad P_1 A_1 = \tilde{L}_1 R_1, \quad A_2 = R_1 P_1^T \tilde{L}_1 = \tilde{L}_1^{-1} (P_1 A P_1^T) \tilde{L}_1. \quad (13.7)$$

Численный пример

14. В качестве простого численного примера модифицированного процесса мы взяли матрицу третьего порядка, которую Рутисхаузер (1958) использовал для иллюстрации случая, когда обычный процесс не сходится. Матрица A_1 , ее собственные значения и матрицы X и Y равны

$$A_1 = \begin{bmatrix} 1 & -1 & 1 \\ 4 & 6 & -1 \\ 4 & 4 & 1 \end{bmatrix}, \quad \begin{matrix} \lambda_1 = 5, \\ \lambda_2 = 2, \\ \lambda_3 = 1, \end{matrix}$$

$$X = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}. \quad (14.1)$$

Ведущий главный минор X первого порядка равен нулю, и, следовательно, мы не можем быть уверены, что обычный процесс сходится, а если он сходится, то собственные значения будут расположены в правильном порядке. Элементы T_s в этом случае расходятся, что легко проверить, используя результаты § 5.

В табл. 1 мы привели результаты, полученные на первых трех шагах обычного LR-процесса, из которых ясно видна форма A_s .

Таблица 1

<i>LR-алгоритм</i>		
L_1	L_2	L_3
$\begin{bmatrix} 1 & & \\ 4 & 1 & \\ 4 & 0,8 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & & \\ 20 & 1 & \\ 4 & 0,16 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & & \\ 100 & 1 & \\ 4 & 0,032 & 1 \end{bmatrix}$
A_2	A_3	A_4
$\begin{bmatrix} 1 & -0,2 & 1 \\ 20 & 6 & -5 \\ 4 & 0,8 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & -0,04 & 1 \\ 100 & 6 & -25 \\ 4 & 0,16 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & -0,008 & 1 \\ 500 & 6 & -125 \\ 4 & 0,032 & 1 \end{bmatrix}$

LR-алгоритм с перестановками:

1-й шаг полностью

$$1' = 2; N_1 = \begin{bmatrix} 1 & & \\ -0,25 & 1 & \\ -1 & 0 & 1 \end{bmatrix}; \quad 2' = 2; N_2 = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & -0,8 & 1 \end{bmatrix};$$

$$R_1 = \begin{bmatrix} 4 & 6 & 1 \\ & -2,5 & 1,25 \\ & & 1 \end{bmatrix}; \quad A_2 = R_1 I_{12} N_1^{-1} N_2^{-1} = \begin{bmatrix} 6 & 3,2 & -1 \\ -1,25 & 1 & 1,25 \\ 1 & 0,8 & 1 \end{bmatrix}$$

2-й и 3-й шаги

$$A_3 = \begin{bmatrix} 5,167 & 3,040 & -1,000 \\ -0,174 & 1,833 & 1,042 \\ 0,167 & 0,160 & 1,000 \end{bmatrix} \quad A_4 = \begin{bmatrix} 5,032 & 3,008 & -1,000 \\ -0,033 & 1,968 & 1,008 \\ 0,032 & 0,032 & 1,000 \end{bmatrix}$$

$$A_\infty = \begin{bmatrix} 5 & 3 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Табл. 1 дает также результаты, полученные при использовании трех шагов модифицированного процесса. Первый шаг показан детально. При приведении A_1 к треугольному виду первая строка переставлена со второй. Это единственная перестановка. Следовательно, для получения A_2 из R_1 мы сначала переставляем первый и второй столбцы и затем последовательно умножаем на N_1^{-1} и N_2^{-1} . На следующих шагах перестановки не необходимы, так что мы используем обычный LR -алгоритм. Матрицы A_s сходятся весьма быстро к верхней треугольной, и легко видеть, что A_∞ это матрица, данная в конце таблицы.

Использование перестановок обеспечило не только сходимость, но и размещение собственных значений на диагонали предельной матрицы в убывающем порядке.

Сходимость модифицированного процесса

15. Доказательство, которое мы дали для сходимости LR -алгоритма, не годится для модифицированного процесса. Однако если сходимость к верхней треугольной форме *действительно* имеет место, и если ни одно из собственных значений не равно нулю, то очевидно, что перестановки в конечном счете должны быть прекращены, так как поддиагональные элементы стремятся к нулю. Это проиллюстрировано в примере § 14.

К несчастью, легко построить простые примеры, когда обычный LR -процесс сходится, а модифицированный не сходится. Рассмотрим, например, матрицу

$$A_1 = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}, \quad \lambda_1 = 3, \quad \lambda_2 = -2. \quad (15.1)$$

Имеем

$$A_2 = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix} = A_1, \quad (15.2)$$

и перестановки нужны как при приведении A_1 , так и A_2 . Так как $A_3 = A_1$, процесс заиклился, и сходимости нет. С другой стороны, обычный LR -алгоритм сходится и дает собственные значения в правильном порядке.

Мы вновь вернемся к этой проблеме в § 24 и покажем, что этот пример несходимости не столь неприятен, как кажется. Пока ограничимся замечанием, что, по нашему мнению, требование численной устойчивости является первостепенным, и поэтому следует пользоваться модифицированным процессом.

Предварительная подготовка исходной матрицы

16. В связи со вторым замечанием § 12 может показаться, что объем работы при использовании LR -алгоритма чрезмерно велик, если матрица A_1 имеет мало нулевых элементов. Однако если A_1 — матрица некоторого частного вида, *инвариантного по отношению к LR -алгоритму*, то объем работы может быть значительно уменьшен.

Этому требованию удовлетворяют два главных вида матриц. Первый — это матрицы в верхней форме Хессенберга, которые инвариантны по отношению к модифицированному LR -алгоритму (и следовательно, по отношению и к обычному LR -алгоритму). Второй состоит из ленточных матриц, ненулевые элементы которых располагаются в полосе, симметричной

относительно главной диагонали. Такие матрицы инвариантны по отношению к обычному LR-алгоритму и не инвариантны по отношению к модифицированному (см. §§ 59—68). Наш опыт показал, что LR-алгоритм практически полезен лишь для матриц этих двух видов. Ввиду того, что существует несколько устойчивых методов приведения матрицы к верхней форме Хессенберга, это не является сильным ограничением.

Инвариантность верхней формы Хессенберга

17. Сначала докажем, что верхняя форма Хессенберга инвариантна по отношению к модифицированному алгоритму. Мы уже обсудили приведение верхних матриц Хессенберга к треугольному виду при помощи метода Гаусса с перестановками (гл. 4, § 33). На r -м шаге приведения выбор ведущего элемента возможен только из строк r и $r+1$, и единственный отличный от нуля элемент соответствующей матрицы N_r (за исключением единичных диагональных элементов) находится в позиции $(r+1, r)$. Матрица N_r поэтому имеет вид $M_{r+1, r}$ (гл. 1, § 40 (vii)). Нужно показать, что матрица A_2 вида

$$A_2 = R_1 I'_{12} M_{21}^{-1} I'_{23} M_{32}^{-1} \dots I'_{n-1, n} M_{n, n-1}^{-1} \quad (17.1)$$

есть верхняя матрица Хессенберга. Здесь $I'_{r, r+1}$ обозначает либо $I_{r, r+1}$, либо I , в зависимости от того, необходима или нет перестановка на r -м шаге приведения к треугольному виду.

Будем доказывать по индукции. Предположим, что матрица

$$R I'_{23} M_{21}^{-1} I'_{34} M_{43}^{-1} \dots I'_{r-1, r} M_{r, r-1}^{-1}$$

имеет первые $r-1$ столбцов в верхней форме Хессенберга, а остальные столбцы в треугольной форме, так что, например, при $n=6$, $r=4$ она имеет вид

$$\begin{matrix} r-1 \\ n-r+1 \end{matrix} \left\{ \begin{array}{ccc|ccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ \hline & & \times & \times & \times & \times \\ & & & \times & \times & \\ & & & & \times & \times \\ & & & & & \times \end{array} \right\}. \quad (17.2)$$

Следующий шаг — умножение справа на $I'_{r, r+1}$. Если этот множитель $I_{r, r+1}$, мы переставляем столбцы r и $r+1$, в противном случае он ничего не меняет. В результате получается матрица вида (a) или (b):

$$(a) = \begin{matrix} r-1 \\ n-r+1 \end{matrix} \left\{ \begin{array}{ccc|ccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ \hline & & \times & \times & \times & \times \\ & & & \times & 0 & \times \\ & & & & \times & \times \\ & & & & & \times \end{array} \right\}, \quad (b) = \begin{matrix} r-1 \\ n-r+1 \end{matrix} \left\{ \begin{array}{ccc|ccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ \hline & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times \end{array} \right\}. \quad (17.3)$$

Умножение справа на $M_{r+1, r}^{-1}$ состоит в добавлении к столбцу r столбца $r+1$, умноженного на множитель, и, следовательно, в результате

получается матрица вида (c) или (d):

$$(c) = \begin{matrix} r \\ \left[\begin{array}{cccc|cc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ \hline & & & \times & \times & \times \\ & & & & \times & 0 \\ & & & & & \times \end{array} \right] \end{matrix}, \quad (d) = \begin{matrix} r \\ \left[\begin{array}{cccc|cc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ \hline & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times \end{array} \right] \end{matrix}. \quad (17.4)$$

В обоих случаях матрица имеет первые r столбцов в верхней форме Хессенберга, а остальные столбцы в треугольной форме. Нулевой элемент в (c) не имеет особого значения. Результат установлен.

Для приведения к треугольному виду нужно приблизительно $n^2/2$ умножений, и еще $n^2/2$ для последующего умножения справа, так что для одного полного шага LR -процесса производится n^2 умножений по сравнению с $2n^3/3$ для полной матрицы.

Одновременное действие со строками и столбцами

18. Дальнейшее преимущество формы Хессенберга состоит в том, что разложение A_s и дальнейшее умножение могут быть совмещены таким образом, что на вычислительной машине с двумя типами памяти s -я итерация будет требовать лишь одного переноса A_s из внешней памяти и ее замены на A_{s+1} . С этой целью необходимо запоминать A_s по столбцам.

Заметим, что для разложения требуется только $n - 1$ множителей $m_{r+1,r}$ ($r = 1, \dots, n - 1$) и для определения первых r из этих множителей нам потребуются только первые r столбцов матрицы A_s . Мы можем описать предлагаемую схему, рассмотрев типичный основной шаг. К началу этого шага вычислены первые $r - 1$ столбцов матрицы A_{s+1} , и записаны на места соответствующих столбцов матрицы A_s во внешней памяти. Вычисление r -го столбца A_{s+1} еще не закончено, и частично обработанный столбец находится в быстродействующей памяти. Элементы $m_{21}, \dots, m_{r,r-1}$ и информация о произведенных перестановках на шагах, в которых эти элементы вычислялись, тоже находятся в быстродействующей памяти. Типичный шаг может быть описан следующим образом:

(i) Переносим $(r + 1)$ -й столбец в быстродействующую память. Для всех значений i от 1 до r проделываем шаги (ii) и (iii).

(ii) Меняем местами $a_{i,r+1}$ и $a_{i+1,r+1}$ в зависимости от того, была ли перестановка при вычислении $m_{i+1,i}$.

(iii) Вычисляем $a_{i+1,r+1} - m_{i+1,i}a_{i,r+1}$ и записываем на место $a_{i+1,r+1}$. По завершении (ii) и (iii) столбец $r + 1$ подвергся всем операциям, входящим в приведение первых r строк A_s .

(iv) Если $|a_{r+2,r+1}| > |a_{r+1,r+1}|$, переставляем эти два элемента и затем вычисляем $m_{r+2,r+1} = a_{r+2,r+1}/a_{r+1,r+1}$. Запоминаем $m_{r+2,r+1}$ и регистрируем перестановку, если она имела место. Заменяем $a_{r+2,r+1}$ на нуль.

Столбец $r + 1$ теперь становится $(r + 1)$ -м столбцом треугольной матрицы, которая была бы найдена после выполнения обычного приведения.

(v) Меняем местами модифицированные столбцы r и $r + 1$ (которые находятся в быстродействующей памяти), если перед вычислением $m_{r+1,r}$ были перестановки.

(vi) К столбцу r прибавляем $(r + 1)$ -й столбец, умноженный на $m_{r+1, r}$. Полученный r -й столбец является теперь r -м столбцом A_{s+1} и может быть помещен во внешнюю память.

Комбинированный процесс дает результат одинаковый с первоначальным процессом даже в смысле ошибок округления и не требует больше арифметических действий. Заметим, что мы можем предположить, что ни один из поддиагональных элементов $a_{r+1, r}$ матрицы A_s не равен нулю, ибо в противном случае мы имели бы дело с двумя или более меньшими матрицами Хессенберга. Поэтому в шаге (iv) знаменатель $a_{r+1, r+1}$ никогда не равен нулю, так как на стадии, когда производится деление, $|a_{r+1, r+1}|$ больше заданного $|a_{r+2, r+1}|$ и некоторых других чисел.

Ускорение сходимости

19. Несмотря на то, что предварительное приведение к форме Хессенберга очень существенно уменьшает объем вычислений, модифицированный LR-алгоритм все еще неэкономичен по сравнению с прежними методами, если только мы не сможем улучшить скорость сходимости.

Мы видели, что для матриц общего вида элемент в позиции (i, j) ($i > j$) стремится к нулю приблизительно как $(\lambda_i/\lambda_j)^2$. Для матриц Хессенберга единственные отличные от нуля поддиагональные элементы находятся в позициях $(r + 1, r)$. Рассмотрим матрицу $(A - pI)$; ее собственные значения равны $(\lambda_i - p)$, и по крайней мере для обычного LR-алгоритма последовательность $a_{n, n-1}^{(s)}$ стремится к нулю как $\{(\lambda_n - p)/(\lambda_{n-1} - p)\}^2$. Если p является хорошим приближением к λ_n , то элемент $a_{n, n-1}^{(s)}$ будет убывать быстро. Поэтому может случиться, что лучше работать с $(A_s - pI)$ вместо A_s .

Заметим, в частности, что если p в точности равен λ_n , и если осуществлять модифицированный процесс, используя точные вычисления, то $(n, n - 1)$ — элемент будет нулем после одной итерации. Действительно, рассмотрим процесс приведения к треугольному виду. Ни один из ведущих элементов (диагональные элементы треугольной матрицы R) не может быть нулем, за исключением последнего, потому что на каждой стадии приведения ведущий элемент равен $a_{r+1, r}$ или некоторому другому числу, а мы предполагали, что первоначально ни один из $a_{r+1, r}$ не равен нулю. Следовательно, так как определитель $(A - \lambda_n I)$ равен нулю, последний ведущий элемент должен быть равен нулю, и треугольная матрица R имеет вид

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & \times & \times \\ & & & & 0 \end{bmatrix}, \quad (19.1)$$

так что вся ее последняя строка состоит из нулевых элементов. Последующее умножение просто комбинирует столбцы, и, следовательно, в конце итерации матрица имеет вид

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \\ & & & 0 & 0 \end{bmatrix}. \quad (19.2)$$

Объединение сдвигов

20. Обсуждение предыдущего параграфа приводит к следующей модификации LR -алгоритма. На s -м шаге осуществляется разложение в произведение треугольных матриц $(A_s - k_s I)$ (вместо матрицы A_s), где k_s — некоторое подходящее число. Соответственно составляется последовательность матриц вида

$$A_s - k_s I = L_s R_s, \quad R_s L_s + k_s I = A_{s+1}. \quad (20.1)$$

Имеем

$$A_{s+1} = R_s L_s + k_s I = L_s^{-1} (A_s - k_s I) L_s + k_s I = L_s^{-1} A_s L_s, \quad (20.2)$$

и, следовательно, матрицы A_s снова подобны A_1 . Действительно,

$$\begin{aligned} A_{s+1} &= L_s^{-1} A_s L_s = L_s^{-1} L_{s-1}^{-1} A_{s-1} L_{s-1} L_s = \\ &= L_s^{-1} \dots L_2^{-1} L_1^{-1} A_1 L_1 L_2 \dots L_s \end{aligned} \quad (20.3)$$

или

$$L_1 L_2 \dots L_s A_{s+1} = A_1 L_1 L_2 \dots L_s. \quad (20.4)$$

Эта модификация обычно называется LR со сдвигом и «восстановлением», так как сдвиг *прибавляется снова* на каждом этапе. Иногда для программирования удобнее использовать «не восстанавливающийся» процесс,

$$A_s - z_s I = L_s R_s, \quad R_s L_s = A_{s+1}, \quad (20.5)$$

так что

$$\begin{aligned} A_{s+1} &= R_s L_s = L_s^{-1} (A_s - z_s I) L_s = L_s^{-1} A_s L_s - z_s I = \\ &= L_s^{-1} (L_{s-1}^{-1} A_{s-1} L_{s-1} - z_{s-1} I) L_s - z_s I = \\ &= L_s^{-1} L_{s-1}^{-1} A_{s-1} L_{s-1} L_s - (z_{s-1} + z_s) I. \end{aligned} \quad (20.6)$$

Продолжая так же, находим

$$\begin{aligned} A_{s+1} &= L_s^{-1} \dots L_2^{-1} L_1^{-1} A_1 L_1 L_2 \dots L_s - (z_1 + z_2 + \dots + z_s) I = \\ &= L_s^{-1} \dots L_2^{-1} L_1^{-1} [A_1 - (z_1 + z_2 + \dots + z_s) I] L_1 L_2 \dots L_s, \end{aligned} \quad (20.7)$$

так что собственные значения A_{s+1} отличаются от собственных значений A_1 на $\sum_{i=1}^s z_i$.

Возвращаясь к восстанавливающемуся варианту, получим

$$\begin{aligned} L_1 L_2 \dots L_{s-1} (L_s R_s) R_{s-1} \dots R_2 R_1 &= \\ &= L_1 L_2 \dots L_{s-1} (A_s - k_s I) R_{s-1} \dots R_2 R_1 = \\ &= (A_1 - k_s I) L_1 L_2 \dots L_{s-1} R_{s-1} \dots R_2 R_1 = \\ &= (A_1 - k_s I) (A_1 - k_{s-1} I) L_1 L_2 \dots L_{s-2} R_{s-2} \dots R_2 R_1 = \\ &= (A_1 - k_s I) (A_1 - k_{s-1} I) \dots (A_1 - k_1 I). \end{aligned} \quad (20.8)$$

Следовательно, если

$$T_s = L_1 L_2 \dots L_s, \quad U_s = R_s \dots R_2 R_1, \quad (20.9)$$

то $T_s U_s$ есть разложение в произведение треугольных матриц $\prod_{i=1}^s (A_1 - k_i I)$, причем порядок сомножителей несуществен. Этот результат нам понадобится позже.

Можно комбинировать сдвиги с перестановками. Каждый член в такой последовательности матриц подобен A_1 , но здесь не имеется какой-либо простой аналогии уравнений (20.8).

Выбор сдвига

21. На практике мы сталкиваемся с проблемой выбора последовательности k_s , которая давала бы быструю сходимость. Если модули всех собственных значений различны, то, независимо от того, вещественные они или комплексные, мы ожидаем, что $a_{n,n-1}^{(s)}$ и $a_{nn}^{(s)}$ стремятся к нулю и к λ_n соответственно. Следовательно, можно взять $k_s = a_{nn}^{(s)}$, если $a_{n,n-1}^{(s)}$ начинает становиться «малым» или $a_{nn}^{(s)}$ начинает стабилизироваться. На самом деле легко показать, что если $a_{n,n-1}^{(s)}$ порядка ε , и мы возьмем $k_s = a_{nn}^{(s)}$, то имеем

$$a_{n,n-1}^{(s+1)} = O(\varepsilon^2). \quad (21.1)$$

При этом, например, для $n = 6$ матрица $(A_s - a_{nn}^{(s)} I)$ будет иметь вид (a) из (21.2):

$$(a) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \varepsilon & 0 \end{bmatrix}, \quad (b) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & a & b \\ & & & & \varepsilon & 0 \end{bmatrix},$$

$$(c) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & a & b \\ & & & & & \bullet - be/a \end{bmatrix}. \quad (21.2)$$

Рассмотрим теперь приведение $(A_s - a_{nn}^{(s)} I)$ к треугольному виду по методу Гаусса с перестановками. Когда приведение почти закончено и осталось привести лишь последнюю строку, текущая матрица имеет вид (b) из (21.2). Элемент a в $(n-1)$ -й строке не будет малым, если только $a_{nn}^{(s)}$ случайно не будет как-либо специально связан с ведущим главным минором порядка $(n-1)$ матрицы A , например, если сдвиг является собственным значением этого минора. Следовательно, на последнем шаге не потребуются перестановка, и мы имеем

$$m_{n,n-1} = \varepsilon/a, \quad (21.3)$$

и треугольная матрица поэтому будет иметь вид (c) из (21.2).

Рассмотрим теперь правостороннее умножение. При умножении справа на все множители, кроме $I'_{n-1, n} M_{n, n-1}^{-1}$, текущая матрица будет иметь вид (a) или (b) из (21.4)

$$(a) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & a & 0 & b \\ & & & & b\epsilon & \\ & & & & -\frac{b\epsilon}{a} & \end{bmatrix}, \quad (b) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & a \\ & & & & & b \\ & & & & & -\frac{b\epsilon}{a} \end{bmatrix} \quad (21.4)$$

в зависимости от того, делалась или нет перестановка перед вычислением $m_{n-1, n-2}$ при приведении. Чтобы закончить умножение справа, мы прибавляем к $(n-1)$ -му столбцу n -й, умноженный на $m_{n, n-1}$, и не делаем никаких перестановок. Поэтому окончательно матрица будет иметь вид (a) или (b) из (21.5)

$$(a) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & b\epsilon/a & b \\ & & & & -b\epsilon^2/a^2 & -b\epsilon/a \end{bmatrix}, \quad (b) \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & (a + b\epsilon/a) & b \\ & & & & -b\epsilon^2/a^2 & -b\epsilon/a \end{bmatrix} \quad (21.5)$$

Следовательно, после восстановления сдвига имеем

$$a_{nn}^{(s+1)} = a_{nn}^{(s)} - b\epsilon/a, \quad a_{n, n-1}^{(s+1)} = -b\epsilon^2/a^2, \quad (21.6)$$

так что $a_{nn}^{(s)}$, конечно, сходится, и $a_{n, n-1}^{(s+1)}$ порядка ϵ^2 , что мы и хотели показать. Заметим, что любые перестановки, которые могут иметь место на других шагах приведения, играют малую роль.

Исчерпывание матрицы

22. Вообще говоря, если $a_{n, n-1}^{(s)}$ стал малым, он будет быстро уменьшаться. Начиная с некоторого места, в соответствии с точностью вычислений, его можно рассматривать как нуль, и тогда текущее значение $a_{nn}^{(s)}$ будет собственным значением. Остальные собственные значения являются собственными значениями ведущей главной подматрицы порядка $(n-1)$. Эта матрица также формы Хессенберга, так что мы можем продолжать тот же процесс, работая с матрицей порядка на единицу меньше исходной. Так как мы ожидаем, что все поддиагональные элементы стремятся к нулю, $a_{n-1, n-2}^{(s)}$ может быть уже достаточно малым, и в этом случае мы в качестве следующего значения k_s можем взять $a_{n-1, n-1}^{(s)}$.

Продолжая таким образом, мы можем по очереди найти все собственные значения, работая с матрицами, порядок которых последовательно уменьшается. Сходимость на последних стадиях к каждому собственному значению будет, вообще говоря, квадратичной и, более того, можно ожидать, что когда предыдущее собственное значение найдено, у нас имеется хорошее начало для отыскания следующего.

Этот метод исчерпывания очень устойчив. По существу, собственные значения могут быть серьезно искажены, только если они чувствительны по отношению к возмущениям поддиагональных элементов. Так как мы предполагаем, что исходная матрица Хессенберга A_1 получена в результате приведения полной матрицы A , то при собственных значениях, чувствительных к малым ошибкам в поддиагональных элементах, мы уже до начала LR -процесса потеряли точность. Единственная дополнительная опасность состоит в том, что обусловленность A_s может последовательно ухудшаться. Мы обсудим это в § 47, когда будем сравнивать LR - и QR -алгоритмы.

Экспериментальная проверка сходимости

23. Мы экспериментировали с LR -алгоритмом (с перестановками) на матрицах Хессенберга, используя сдвиг, определенный текущим значением $a_{nn}^{(s)}$. Естественно, что он использовался лишь для вещественных матриц, про которые известно, что они имеют вещественные собственные значения, или для матриц с комплексными элементами. Вообще говоря, не ясно, когда нужно было бы начинать использовать это значение сдвига, и мы исследовали эффективность следующих процедур:

(i) Сдвиг $a_{nn}^{(s)}$ использовался на каждом этапе.

(ii) Сдвиг использовался, только когда $|1 - a_{nn}^{(s)}/a_{nn}^{(s-1)}| < \varepsilon$, где ε — некоторый допуск (т. е. сдвиг употребляется, когда $a_{nn}^{(s)}$ уже начали стабилизироваться).

(iii) Сдвиг использовался только тогда, когда $|a_{n,n-1}^{(s)}| < \eta \|A_1\|_\infty$, где η — снова некоторый допуск.

Вообще говоря, сходимость при использовании (i) была очень хорошей, и это указывает на то, что при использовании (ii) и (iii) допуски не должны быть слишком малыми. В целом метод (ii) с $\varepsilon = \frac{1}{3}$ был наиболее эффективным. Для большинства испытанных матриц (порядок которых был от 5 до 32) среднее число итераций на каждое собственное значение было порядка 5, и обычно несколько последних собственных значений получались после одной или двух итераций. Однако существует много матриц, которые не дают сходимости при использовании (i), (ii) или (iii). Это, например, матрицы второго порядка

$$\begin{bmatrix} a & b \\ c & 0 \end{bmatrix}, \quad |b|, |c| > |a|, \quad (23.1)$$

частный случай которых уже обсуждался в § 15. При последовательных применениях LR -алгоритма с перестановками, они становятся

$$\begin{bmatrix} a & c \\ b & 0 \end{bmatrix} \quad \text{и} \quad \begin{bmatrix} a & b \\ c & 0 \end{bmatrix}. \quad (23.2)$$

При этом $a_{nn}^{(s)} = 0$ при всех s , и выбор $k_s = a_{nn}^{(s)}$ не дает эффекта. Если мы имеем сходимость, то $a_{n,n-1}^{(s)}$ стремится к нулю, и поэтому использовался сдвиг

$$k_s = \frac{1}{2} a_{n,n-1}^{(s)} + a_{nn}^{(s)}, \quad (23.3)$$

причем $k_s \rightarrow a_{nn}^{(s)}$. Этот простой эмпирический прием давал сходимость почти во всех случаях, когда применение $k_s = a_{nn}^{(s)}$ ничего не давало. Однако существуют матрицы с различными собственными значениями, которые не давали сходимости даже при таком модифицированном сдвиге. В следующем параграфе опишем метод для выбора сдвига, который оказался значительно более эффективным.

Улучшенный выбор сдвига

24. Мотивировкой для улучшения правила выбора сдвига является рассмотрение случая, когда матрица вещественная, но имеет несколько комплексно сопряженных собственных значений. Мы знаем, что в этом случае предельная матрица не будет в точности верхней треугольной. Если собственные значения, имеющие наименьший модуль, суть комплексно сопряженная пара $\lambda \pm i\mu$, то итерации обязательно будут иметь вид (a) из (24.1)

$$(a) \begin{bmatrix} \times & \times & \times & & \times & \times \\ \times & \times & \times & & \times & \times \\ & \times & \times & & \times & \times \\ \hline & & & \varepsilon & \times & \times \\ & & & & \times & \times \end{bmatrix}, \quad (b) \begin{bmatrix} \times & \times & \times & \times & & \times \\ \times & \times & \times & \times & & \times \\ & \times & \times & \times & & \times \\ & & \times & \times & & \times \\ \hline & & & \times & \times & \times \\ & & & & \varepsilon & \times \end{bmatrix}, \quad (24.1)$$

где собственные значения матрицы второго порядка в нижнем правом углу стремятся к соответствующим значениям. С другой стороны, если собственное значение, имеющее наименьший модуль, это вещественное λ , то итерации примут вид (b) из (24.1). Предположим, что мы сосчитали собственные значения матрицы второго порядка в обоих случаях. Тогда в случае (a) они стремятся к $\lambda \pm i\mu$, а в случае (b) наименьшее из двух стремится к λ . Установив это, мы на некоторое время оставим случай комплексно сопряженных собственных значений.

Для вещественных матриц с вещественными собственными значениями предлагаем следующий выбор k_s . Вычисляем собственные значения матрицы второго порядка в нижнем правом углу A_s . Если они комплексные и равны $p_s \pm iq_s$, берем $k_s = p_s$. Если они вещественные и равны p_s и q_s , берем k_s равным p_s или q_s в зависимости от того, что меньше: $|p_s - a_{nn}^{(s)}|$ или $|q_s - a_{nn}^{(s)}|$.

Если исходная матрица комплексная, то, вообще говоря, наша матрица второго порядка будет иметь два комплексных собственных значения p_s и q_s , и снова наш выбор будет зависеть от сравнения $|p_s - a_{nn}^{(s)}|$ и $|q_s - a_{nn}^{(s)}|$.

Мы можем применять сдвиг, определяемый таким образом, на каждом этапе или ждать, пока предложенный сдвиг не даст некоторого указания о сходимости. На практике поведение, по-видимому, не сильно зависит от этого решения. Наилучшие результаты были получены, когда предложенное значение k_s принималось в качестве сдвига, если выполнялось условие

$$|k_s/k_{s-1} - 1| < 1/2. \quad (24.2)$$

Оказалось, что в среднем скорость сходимости при использовании такой техники очень высокая. Этот метод использовался наиболее интенсивно

на матрицах с комплексными элементами. Стоит упомянуть, что такой выбор сдвига для матриц (23.1) дает сходимость в одну итерацию. Так как относящаяся к делу матрица второго порядка теперь полная, это не удивительно, но это наводит на мысль, что такой выбор сдвига может с успехом давать лучшее общее течение процесса.

Комплексно сопряженные собственные значения

25. Возвращаясь теперь к случаю, когда наименьшие собственные значения суть комплексно сопряженная пара, мы можем ожидать, что собственные значения матрицы второго порядка являются приближениями к этой паре. Предположим, что на данном этапе собственные значения этой матрицы равны $p \pm iq$. Если мы выполним два шага обычного LR-алгоритма, используя сначала сдвиг $p - iq$, а потом $p + iq$, то, хотя нам придется, конечно, использовать комплексную арифметику, мы покажем, что при точных вычислениях вторая вычисленная матрица будет вещественной. Обозначая три соответствующие матрицы через A_1 , A_2 и A_3 , имеем

$$\left. \begin{aligned} A_1 - (p - iq)I &= L_1 R_1, & R_1 L_1 + (p - iq)I &= A_2, \\ A_2 - (p + iq)I &= L_2 R_2, & R_2 L_2 + (p + iq)I &= A_3. \end{aligned} \right\} \quad (25.1)$$

Используя (20.8), получим

$$L_1 L_2 R_2 R_1 = [A_1 - (p - iq)I][A_1 - (p + iq)I], \quad (25.2)$$

и очевидно, что правая часть вещественная. Предполагая, что эта вещественная матрица неособенная, т. е. что $p \pm iq$ не являются точными собственными значениями A_1 , получаем, что ее разложение в произведение треугольных, если оно существует, единственно, и треугольные матрицы вещественны. Поэтому треугольная матрица $L_1 L_2$ в (25.2) должна быть вещественной. Так как

$$A_3 = (L_1 L_2)^{-1} A_1 L_1 L_2, \quad (25.3)$$

мы видим, что A_3 также вещественная.

Можно было бы ожидать, что при выполнении двух шагов LR-алгоритма вычисленная матрица A_3 будет иметь пренебрежимо малую мнимую часть, и ее можно считать точно вещественной. К сожалению, это не так, и нет ничего необычного в том, что конечная матрица может иметь весьма значительную мнимую часть. Процесс численно устойчив в том смысле, что собственные значения вычисленной A_3 близки к собственным значениям A_1 (если только A_1 неплохо обусловлена по отношению к проблеме собственных значений), но так же, как часто оказывалось раньше, *нет никаких гарантий, что окончательно вычисленная матрица будет близка к матрице, которая получилась бы при точных вычислениях.*

26. Для того чтобы увидеть, как это может получиться, рассмотрим случай, когда A_1 есть квадратная матрица второго порядка из (26.1) с собственными значениями λ и $\bar{\lambda}$. Имеем

$$A_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A_2 = \begin{bmatrix} \bar{\lambda} & b \\ 0 & \lambda \end{bmatrix}, \quad (A_2 - \bar{\lambda}I) = \begin{bmatrix} 0 & b \\ 0 & \lambda - \bar{\lambda} \end{bmatrix}. \quad (26.1)$$

Если выполним один шаг LR-алгоритма, используя сдвиг, в точности равный λ , то получим выписанную в (26.1) матрицу A_2 , и первый столбец $(A_2 - \bar{\lambda}I)$ состоит из нулей. Следовательно, разложение в произведение

треугольных $(A_2 - \bar{\lambda}I)$ не единственно. Соответственно, если применим сдвиги $(\lambda + \varepsilon)$ и $(\bar{\lambda} + \bar{\varepsilon})$, где ε мало, то получим

$$A_2 - (\bar{\lambda} + \bar{\varepsilon})I = \begin{bmatrix} \varepsilon_1 & b \\ \varepsilon_2 & (\lambda - \bar{\lambda}) + \varepsilon_3 \end{bmatrix}, \quad (26.2)$$

где ε_1 , ε_2 и ε_3 того же порядка, что и ε . Если вычисления проводятся точно, то A_3 будет вещественной, но малые ошибки ε_1 и ε_2 существенно будут менять A_3 . При вычислениях с t знаками A_3 будет, вообще говоря, иметь мнимые компоненты порядка $2^{-t}/\varepsilon$, и ими нельзя пренебрегать. При этом собственные значения полностью сохраняются, если удерживать мнимые компоненты. Здесь снова полезно рассмотреть случай, когда используются точные значения λ и $\bar{\lambda}$. В разложении в произведение треугольных матриц матрицы $(A_2 - \bar{\lambda}I)$ в (26.1) в качестве L_2 можно взять

$$\begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix}, \quad (26.3)$$

где m — произвольное. Окончательная матрица A_3 имеет вид

$$\begin{bmatrix} \bar{\lambda} + mb & b \\ m[(\lambda - \bar{\lambda}) - mb] & \lambda - mb \end{bmatrix}. \quad (26.4)$$

так что A_3 в точности подобна A_1 при всех значениях m , но она не вещественная, если только m не выбрано так, что $\lambda - mb$ вещественно. Неопределенность при точных вычислениях со сдвигом, определенным точными собственными значениями, связана с плохой определенностью m в практических процессах. Это иллюстрируется в табл. 2.

Таблица 2

$$A_1 = \begin{bmatrix} 0,31257 & 0,61425 \\ -0,51773 & 0,41631 \end{bmatrix} \quad \begin{array}{l} \text{Собственные значения } 0,36444 \pm 0,56154i \\ \text{Сдвиги } \lambda, \bar{\lambda} = 0,36442 \pm 0,56156i \end{array}$$

$$\bar{A}_2 - \bar{\lambda}I = \begin{bmatrix} 0,00000 + 0,00004i & 0,61425 \\ -0,00003 - 0,00004i & 0,00004 + 1,12308i \end{bmatrix}$$

$$\bar{A}_3 = \begin{bmatrix} -0,24983 - 0,10083i & 0,61425 \\ -1,11108 - 0,20167i & 0,97871 + 0,10083i \end{bmatrix}$$

Собственные значения A_1 равны $0,36444 \pm 0,56154i$, и использовались сдвиги $0,36442 \pm 0,56156i$. Оба элемента в первом столбце $(\bar{A}_2 - \bar{\lambda}I)$ очень малы и имеют очень большие относительные ошибки. Матрица A_3 , которая должна бы быть вещественной, если бы вычисления выполнялись точно, имеет мнимую часть того же порядка, что и вещественная. Тем не менее ее собственные значения совпадают с рабочей точностью с собственными значениями A_1 . Кроме того, если, вычислив собственные векторы $\bar{A}_3$, мы используем их для получения собственных векторов A_1 , применяя вычисленные преобразования, то полученные векторы будут точными собственными векторами A_1 .

Критика модифицированного LR-алгоритма

27. Обсуждая комплексно сопряженные сдвиги, мы опустили вопрос перестановок. Точные вычисления не гарантируют, вообще говоря, что A_3 будет вещественной, если мы включим перестановки. Однако если мы будем использовать перестановки только при переходе от A_1 к A_2 , то

в обозначениях § 13 имеем

$$\left. \begin{aligned} P_1(A_1 - \lambda I) &= \tilde{L}_1 R, & R_1 P_1^T \tilde{L}_1 + \lambda I &= A_2, & A_2 &= \tilde{L}_1^{-1} (P_1 A_1 P_1^T) \tilde{L}_1, \\ A_2 - \bar{\lambda} I &= L_2 R_2, & R_2 L_2 + \bar{\lambda} I &= A_3, & A_3 &= L_2^{-1} A_2 L_2, \end{aligned} \right\} \quad (27.1)$$

что дает

$$\begin{aligned} \tilde{L}_1 L_2 R_2 R_1^* &= \tilde{L}_1 (A_2 - \bar{\lambda} I) R_1 = (P_1 A_1 P_1^T - \bar{\lambda} I) \tilde{L}_1 R_1 = \\ &= (P_1 A_1 P_1^T - \bar{\lambda} I) P_1 (A_1 - \lambda I), \end{aligned} \quad (27.2)$$

и умножение справа на P^T показывает, что последнее произведение вещественное. К сожалению, мы не можем гарантировать, что перестановки не требуются при переходе от A_2 к A_3 .

Даже если мы ограничимся случаем, когда нет комплексно сопряженных собственных значений (а ситуации, когда это можно гарантировать а priori, редки), модифицированный алгоритм все же может быть подвергнут критике с точки зрения численной устойчивости. Как мы установили в главе 4, устойчивость разложения в произведение треугольных даже с использованием перестановок не может быть безусловно гарантирована. При предположении, что потребуется не более четырех итераций на одно собственное значение, число итераций при работе с матрицей 50-го порядка равно примерно 200. Нельзя пренебрегать опасностью роста наддиагональных элементов. (Диагональные элементы стремятся к собственным значениям, а поддиагональные к нулю). Опыт показывает, что потеря точности, вызванная этой причиной, довольно обычна.

QR-алгоритм

28. В соответствии с общей точкой зрения этой книги естественно попытаться заменить устойчивые элементарные преобразования в модифицированном LR-алгоритме элементарными унитарными преобразованиями. Это прямо приводит к QR-алгоритму Френсиса — Кублановской (1964). Вместо разложения в произведение треугольных здесь используется разложение в произведение унитарной матрицы Q и верхней треугольной R .

Алгоритм определяется соотношениями

$$A_{s+1} = Q_s R_s, \quad A_{s+1} = Q_s^H A_s Q_s = Q_s^H Q_s R_s Q_s = R_s Q_s, \quad (28.1)$$

так что на каждой стадии теперь используем унитарное преобразование подобия. Требуемая факторизация уже была обсуждена в гл. 4, §§ 46—55, и мы показали, что если A_s неособенная, то эта факторизация по существу единственна, и она безусловно единственна, если мы возьмем диагональные элементы R_s вещественными и положительными. Если A_s — вещественная, то Q_s и R_s — вещественные. Эта факторизация имеет то преимущество, что обращение в нуль ведущего главного минора A_s не вызывает нарушения алгоритма, как это было в обычном LR-разложении.

Последовательные итерации удовлетворяют соотношениям, похожим на соотношения, выведенные для LR-преобразования. Имеем

$$A_{s+1} = Q_s^H A_s Q_s = Q_s^H Q_{s-1}^H A_{s-1} Q_{s-1} Q_s = Q_s^H \dots Q_2^H Q_1^H A_1 Q_1 Q_2 \dots Q_s, \quad (28.2)$$

что дает

$$Q_1 Q_2 \dots Q_s A_{s+1} = A_1 Q_1 Q_2 \dots Q_s. \quad (28.3)$$

Поэтому все A_s унитарно подобны A_1 , и если положить

$$Q_1 Q_2 \dots Q_s = P_s, \quad R_s R_{s-1} \dots R_1 = U_s, \quad (28.4)$$

то

$$\begin{aligned} P_s U_s &= Q_1 \dots Q_{s-1} (Q_s R_s) R_{s-1} \dots R_1 = Q_1 \dots Q_{s-1} A_s R_{s-1} \dots R_1 = \\ &= A_1 Q_1 \dots Q_{s-1} R_{s-1} \dots R_1 = A_1 P_{s-1} U_{s-1}. \end{aligned} \quad (28.5)$$

Следовательно,

$$P_s U_s = A_1^s, \quad (28.6)$$

так что P_s и U_s дают соответствующую факторизацию A_1^s . Эта факторизация единственна, если диагональные элементы верхней треугольной матрицы положительны, и так будет для U_s , если это верно для всех R_s .

Сходимость QR-алгоритма

29. Вообще говоря, A_s стремятся к верхней треугольной форме при несколько менее ограничительных условиях, необходимых для сходимости LR-алгоритма. Прежде чем детально рассматривать сходимость, укажем на одну небольшую разницу между сходимостями этих двух методов.

Если A_1 — верхняя треугольная, то в LR-алгоритме мы имеем $L_1 = I$, $R_1 = A_1$ и, следовательно, $A_s = A_1$ при всех s . Это не так в QR-алгоритме, если нам нужно, чтобы R_s имела положительные диагональные элементы. В самом деле, положив

$$a_{ii} = |a_{ii}| \exp(i\theta_i), \quad D = \text{diag}[\exp(i\theta_i)], \quad (29.1)$$

имеем

$$A_1 = D (D^{-1} A_1), \quad A_2 = D^{-1} A_1 D. \quad (29.2)$$

Следовательно, хотя A_2 имеет те же самые диагональные элементы, что и A_1 , наддиагональные элементы умножаются на комплексный множитель, равный по модулю единице. Поэтому очевидно, что мы не можем ожидать, что матрицы A_s стремятся к пределу в буквальном смысле, если только не все собственные значения вещественны и положительны. Множитель, по модулю равный единице, несуществен, и мы будем говорить, что матрицы A_s «в основном» сходятся, если асимптотически $A_{s+1} \sim D^{-1} A_s D$ при некоторой унитарной диагональной D .

Как в § 5, мы сначала рассмотрим случай матрицы третьего порядка с

$$|\lambda_1| > |\lambda_2| > |\lambda_3|,$$

так что A_1^s имеет вид (5.5). Факторизацию удобно осуществлять при помощи процесса ортогонализации Шмидта с нормировкой (гл. 4, § 54). Первый столбец P_s поэтому равен первому столбцу A_1^s , нормированному так, что его евклидова норма равна единице. Его компоненты находятся в отношениях

$$\lambda_1^s x_1 a_1 + \lambda_2^s y_1 a_2 + \lambda_3^s z_1 a_3 : \lambda_1^s x_2 a_1 + \lambda_2^s y_2 a_2 + \lambda_3^s z_2 a_3 : \lambda_1^s x_3 a_1 + \lambda_2^s y_3 a_2 + \lambda_3^s z_3 a_3. \quad (29.3)$$

Если предположить, что a_1 не равен нулю, то эти элементы обязательно будут в отношении $x_1 : x_2 : x_3$. Если считать, что все диагональные элементы всех R_i должны быть положительными, то первый столбец P_s

принимает асимптотически вид

$$(a_1/|a_1|)(\lambda_1/|\lambda_1|)^s(x_1, y_1, z_1)/(|x_1|^2 + |y_1|^2 + |z_1|^2)^{1/2}, \quad (29.4)$$

что дает сходимость в основном к первому столбцу унитарной матрицы, полученной при факторизации X . Если a_1 равен нулю, но a_2 — нет, то компоненты первого столбца P_s будут находиться в отношении $y_1 : y_2 : y_3$. Подобное явление встречается в связи с LR-техникой.

Заметим, что обращение в нуль x_1 безразлично для QR-алгоритма, в то время как это вызывает расходимость LR-алгоритма. «Случайный» срыв LR-процесса, вызванный тем, что $\lambda_1^s x_1 a_1 + \lambda_2^s y_1 a_2 + \lambda_3^s z_1 a_3$ при некотором s обратилось в нуль, не имеет аналога в QR-технике.

Мы могли бы продолжать наш элементарный анализ, чтобы выяснить асимптотическое поведение второго и третьего столбцов P_s , но это значительно более трудно, чем в LR-алгоритме. Ясно, что мы можем ожидать, что P_s будут в основном сходиться к унитарному сомножителю X .

Формальное доказательство сходимости

30. Сейчас мы дадим формальное доказательство сходимости в основном P_s . Сначала предположим, что собственные значения A таковы, что

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|. \quad (30.1)$$

Тогда A_1 необходимо имеет линейные элементарные делители, и мы можем написать

$$A_1^s = X \operatorname{diag}(\lambda_i^s) X^{-1} = X D^s Y. \quad (30.2)$$

Определим матрицы Q, R, L, U соотношениями

$$X = QR, \quad Y = LU, \quad (30.3)$$

где R и U — верхние треугольные, L — единичная нижняя треугольная, а Q — унитарная. Все четыре матрицы не зависят от s . То, что R неособенная, следует из неособенности X . Заметим, что QR-разложение всегда существует, а разложение Y в произведение треугольных существует, только если все ее главные ведущие миноры не равны нулю. Мы имеем

$$A_1^s = Q R D^s L U = Q R (D^s L D^{-s}) D^s U, \quad (30.4)$$

и ясно, что $D^s L D^{-s}$ есть единичная нижняя треугольная матрица. Ее (i, j) -элемент равен $l_{ij}(\lambda_i/\lambda_j)^s$ при $i > j$, и следовательно, мы можем написать

$$D^s L D^{-s} = I + E_s^L, \quad \text{где } E_s \rightarrow 0 \text{ при } s \rightarrow \infty. \quad (30.5)$$

Поэтому уравнение (30.4) дает

$$\begin{aligned} A_1^s &= Q R (I + E_s) D^s U = Q (I + R E_s R^{-1}) R D^s U = \\ &= Q (I + F_s) R D^s U; \quad \text{где } F_s \rightarrow 0 \text{ при } s \rightarrow \infty. \end{aligned} \quad (30.6)$$

Матрицу $(I + F_s)$ можно разложить в произведение унитарной $\tilde{Q}_s$ и верхней треугольной $\tilde{R}_s$, и так как $F_s \rightarrow 0$, $\tilde{Q}_s$ и $\tilde{R}_s$ стремятся к I . Следовательно, окончательно имеем

$$A_1^s = (Q \tilde{Q}_s) (\tilde{R}_s R D^s U). \quad (30.7)$$

Первый сомножитель в скобках унитарный, а второй — это верхняя треугольная матрица. При предположении, что A_1^s — неособенная, ее разложение в такое произведение по существу единственно, и поэтому P_s равна $Q\tilde{Q}_s$ с точностью до умножения справа на диагональную унитарную матрицу. Следовательно, P_s сходятся в основном к Q . Если мы хотим, чтобы все R_s имели положительные диагональные элементы, то можем найти унитарный диагональный множитель из (30.7). Написав

$$D = |D| D_1, \quad U = D_2 (D_2^{-1} U), \quad (30.8)$$

где D_1 и D_2 — унитарные диагональные матрицы, и $D_2^{-1} U$ имеет положительные диагональные элементы (R_s и R уже имеют положительные диагональные элементы), мы получим из (30.7)

$$A_1^s = Q\tilde{Q}_s D_2 D_1^s [(D_2 D_1^s)^{-1} \tilde{R}_s R (D_2 D_1^s) |D|^s (D_2^{-1} U)]. \quad (30.9)$$

Матрица в квадратных скобках верхняя треугольная с положительными диагональными элементами, и, следовательно, $P_s \sim Q D_2 D_1^s$, что показывает, что Q_s обязательно сходятся к D_1 .

Нарушение порядка собственных значений

31. Приведенное доказательство показывает, что если мы предположим, что все ведущие главные миноры Y не равны нулю, то получим не только сходимость в основном, но и правильное расположение собственных значений на диагонали. При этом на матрицу X не накладываются соответствующие ограничения. Для матрицы третьего порядка очевидно, что если a_1 (т. е. главный минор первого порядка Y) равен нулю, то P_s все равно стремятся к пределу. Сейчас мы формально установим это в общем случае, когда несколько главных миноров Y равны нулю.

Хотя в этом случае Y не будет иметь разложения в произведение треугольных, всегда существует такая матрица перестановок P , что PY имеет разложение в произведение треугольных. В гл. 4, § 24 мы показали, как определить такую P , используя правило выбора ведущего элемента. В терминах метода Гаусса на каждой стадии берем в качестве ведущего элемента наибольший по модулю элемент в ведущем столбце. Вместо этого предположим, что на r -м этапе мы выбираем в качестве ведущего элемента первый отличный от нуля элемент из (r, r) , $(r+1, r)$, $\dots$, (n, r) , и если это (r', r) -элемент, то переставляем строки в порядке $r', r, r+1, \dots, r'-1, r'+1, \dots, n$. Тогда окончательно имеем

$$PY = LU, \quad (31.1)$$

где L имеет некоторое количество нулей под диагональю в позициях, связанных с матрицей перестановок P . Так как Y — неособенная, то на каждой стадии обязательно будет ненулевой ведущий элемент. Поэтому можно написать

$$A_1^s = X D^s Y = X D^s P^T L U = X P^T (P D^s P^T) L U \quad (31.2)$$

Матрица $P D^s P^T$ диагональная с элементами λ_i^s , расположенными в другом порядке, в то время как $X P^T$ получена из X перестановкой столбцов. Если мы положим

$$X P^T = Q R, \quad P D^s P^T = D_3^s, \quad (31.3)$$

то

$$A_1^s = QRD_3^sLU = QR(D_3^sLD_3^{-s})D_3^sU. \quad (31.4)$$

Так как элементы λ_i^s не находятся в естественном порядке в D_3 , на первый взгляд кажется, что $D_3^sLD_3^{-s}$ уже не стремится к I . Однако если мы обозначим диагональные элементы D_3 через $\lambda_{i_1}, \lambda_{i_2}, \dots, \lambda_{i_n}$, то (p, q) -элемент $(p > q)$ матрицы $D_3^sLD_3^{-s}$ равен $(\lambda_{i_p}/\lambda_{i_q})^s l_{pq}$, и правило выбора ведущего элемента для Y обеспечивает $l_{pq} = 0$, если $i_p < i_q$. Следовательно, мы можем написать $D_3^sLD_3^{-s} = I + E_s$, где $E_s \rightarrow 0$ при $s \rightarrow \infty$. Матрицы P_s поэтому стремятся в основном к матрице Q , полученной при факторизации XP^T .

Равные по модулю собственные значения

32. Рассмотрим случай, когда A_1 имеет несколько равных по модулю собственных значений, но линейные элементарные делители. Будем предполагать, что все ведущие главные миноры Y не равны нулю, так как влияние того, что они обращаются в нуль, уже исследовалось. Имеем

$$A_1^s = XD^sLU. \quad (32.1)$$

Предположим, что $|\lambda_r| = |\lambda_{r+1}| = \dots = |\lambda_t|$, а модули всех остальных собственных значений различны. Элемент D^sLD^{-s} в позиции (i, j) под диагональю равен $(\lambda_i/\lambda_j)^s l_{ij}$, и поэтому стремится к нулю, если только не выполняется неравенство

$$t \geq i > j \geq r, \quad (32.2)$$

и остается по модулю равным l_{ii} в противном случае.

В случае, когда все собственные значения, равные по модулю, равны и алгебраически, можно написать

$$D^sLD^{-s} = \tilde{L} + E_s \quad (E_s \rightarrow 0), \quad (32.3)$$

где $\tilde{L}$ — фиксированная единичная нижняя треугольная матрица, которая равна I , за исключением элементов в позициях (i, j) , равных l_{ij} , где i, j удовлетворяют (32.2). Если положить $XL = QR$, то имеем

$$A_1^s = QR(I + \tilde{L}^{-1}E_s)D^sU = Q(I + R\tilde{L}^{-1}E_sR^{-1})RD^sU = \\ = Q(I + F_s)RD^sU = (Q\tilde{Q}_s)(\tilde{R}_sRD^sU), \quad (F_s \rightarrow 0), \quad (32.4)$$

где $\tilde{Q}_s\tilde{R}_s$ — факторизация $I + F_s$. Следовательно, с точностью до обычного унитарного диагонального сомножителя, P_s стремятся к матрице Q , полученной при факторизации $X\tilde{L}$. Заметим, что столбцы $X\tilde{L}$ являются системой линейно независимых собственных векторов A_1 , так как $X\tilde{L}$ отличается от X только тем, что столбцы от r до t заменены своими линейными комбинациями. Следовательно, кратное собственное значение, которому соответствуют линейные элементарные делители, не препятствует сходимости.

Если равные по модулю собственные значения не равны между собой, то матрица $\tilde{L}$ имеет ту же форму, что и раньше, но ее отличные от нуля поддиагональные элементы меняются от шага к шагу.

Если $\lambda_i = |\lambda_i| \exp(i\theta_i)$, то имеем

$$\tilde{l}_{ij} = l_{ij} \exp[is(\theta_i - \theta_j)]. \quad (32.5)$$

Матрица $X\tilde{L}$ совпадает с X , за исключением столбцов от r до t . Эти столбцы являются линейными комбинациями соответствующих столбцов X . Следовательно, при QR -разложении $X\tilde{L}$ все столбцы Q , за исключением столбцов от r до t , совпадают со столбцами унитарного сомножителя X , а столбцы от r до t являются линейными комбинациями соответствующих столбцов X . Следовательно, с точностью до столбцов от r до t , матрицы P_s в основном сходятся.

Наиболее важным является случай, когда вещественная матрица наряду с вещественными собственными значениями имеет несколько пар комплексно сопряженных собственных значений. Ясно, что, так же как в §§ 9, 10, мы можем показать, что матрицы A_s в основном стремятся к верхней треугольной матрице, за исключением единственных поддиагональных элементов, связанных с каждой парой комплексно сопряженных собственных значений. Каждый такой элемент связан с матрицей второго порядка на диагонали, имеющей собственные значения, которые сходятся к соответствующей комплексно сопряженной паре.

Если мы имеем r комплексных собственных значений, модули которых равны, то в пределе A_s имеет на диагонали матрицу порядка r , собственные значения которой стремятся к этим r значениям. В качестве простого примера рассмотрим матрицу A_1 вида

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \begin{aligned} Q_1 &= A_1, & R_1 &= I, \\ A_2 &= R_1 Q_1 = A, \end{aligned} \quad (32.6)$$

собственные значения которой равны $e^{\frac{1}{2}i\pi r}$ ($r = 0, 1, 2, 3$). Она инвариантна по отношению к QR -алгоритму, и поэтому сама будет единственным диагональным блоком (4-го порядка). Такие блоки трудно различить при автоматической процедуре, особенно если размеры их заранее неизвестны. (См. § 47 о влиянии сдвига.)

Другое доказательство для LR -метода

33. Наметим переложение предыдущего метода для нового доказательства сходимости LR -алгоритма. Если λ_i различны, то имеем

$$A_1^s = X D^s Y = L_X U_X D^s L_Y U_Y, \quad (33.1)$$

где $L_X U_X$ и $L_Y U_Y$ — разложения X и Y в произведение треугольных, для существования которых потребуем, чтобы ведущие главные миноры X и Y не были равными нулю. Следовательно,

$$\begin{aligned} A_1^s &= L_X U_X (D^s L_Y D^{-s}) D^s U_Y = L_X U_X (I + E_s) U_Y = \\ &= L_X (I + F_s) U_X U_Y \quad (E_s \rightarrow 0; F_s \rightarrow 0). \end{aligned} \quad (33.2)$$

Матрица $I + F_s$ наверняка имеет разложение в произведение треугольных при достаточно больших s , причем оба сомножителя стремятся к I , но при малых s это разложение может не существовать. Это соответствует «случайному» срыву, который наступает, когда главный минор A_s при некотором s равен нулю. Это не имеет аналогии в QR -алгоритме. Пренебрегая этой возможностью, видим, что $T_s \rightarrow L_X$ в обозначениях § 4, откуда следует, что матрицы A_s стремятся к верхней треугольной матрице с собственными значениями λ_i , расположенными на диагонали в правильном порядке.

Если один (или более) главный угловой минор Y равен нулю, то тем не менее будет существовать такая матрица перестановок, что PY имеет разложение в произведение треугольных. Обозначая его снова $L_X U_Y$, имеем

$$A_1^s = XD^s P^T L_X U_Y = (XP^T) (PD^s P^T) L_X U_Y. \quad (33.3)$$

Если XP^T имеет разложение $L_X U_X$ в произведение треугольных, т. е. если все ее ведущие главные миноры не равны нулю, то мы обычным образом можем показать, что $T_s \rightarrow L_X$ и A_s стремится к треугольной матрице с диагональю $P D P^T$.

34. Рассмотрение случая кратных собственных значений (с линейными элементарными делителями) и собственных значений с равными модулями аналогично тому, которое было проделано для QR-алгоритма. В первом случае мы имеем сходимость, если $XP^T \tilde{L}$ имеет разложение в произведение треугольных, где $\tilde{L}$ — фиксированная матрица, полученная по нижней треугольной в факторизации Y способом, который мы описали. Однако случай, когда A_1 — симметричная, имеет особый интерес. Мы имеем $X = Y^T$, и если P выбрана так, что PY имеет разложение в произведение треугольных, то XP^T также имеет разложение, так как ведущие главные миноры XP^T и PY совпадают.

Если нет равных по модулю собственных значений, то требуется, чтобы именно XP^T имела бы разложение в произведение треугольных, и на первый взгляд кажется, что LR-процесс должен всегда сходиться. Однако мы не должны забывать, что $(I + F_s)$ может не иметь разложения в произведение треугольных при некоторых специфических значениях s , хотя мы и знаем, что она должна иметь такое разложение при всех достаточно больших s . Следовательно, для симметричных матриц, имеющих различные по модулю собственные значения, точные вычисления должны всегда давать сходимость, независимо от возможности случайного срыва на ранней стадии.

Если A_1 — симметричная и положительно определенная, то мы можем показать, что сходимость должна быть, даже если есть кратные собственные значения. (Так как все собственные значения положительны, то равные по модулю собственные значения обязательно равны.) Легко показать, что в этом случае случайный срыв невозможен ни на какой стадии. Если мы напомним

$$PY = U_1, \quad XP^T = U_1^T L_1^T \quad (34.1)$$

то, как в § 32, мы требуем для сходимости LR-алгоритма, чтобы $XP^T L$ имела разложение в произведение треугольных. Здесь $\tilde{L}$ получена из L_1 вычеркиванием всех элементов под диагональю, за исключением элементов в тех строках и столбцах, которые соответствуют равным диагональным элементам $P D^s P^T$. Мы имеем

$$(XP^T) \tilde{L} = U_1^T L_1^T \tilde{L}, \quad (34.2)$$

и, следовательно, необходимо только, чтобы $L_1^T \tilde{L}$ имела разложение в произведение треугольных, т. е. чтобы ее ведущие главные миноры не были равны нулю. Из связи $\tilde{L}$ и L , используя теорему Бине — Коши, получаем, что все ведущие главные миноры $L_1^T \tilde{L}$ не меньше единицы. Результат установлен.

Практическое применение QR-алгоритма

35. Типичный шаг QR-алгоритма требует выполнения факторизации A_s в QR_s . Для теоретических целей было удобно думать о ней как об ортонормализации Шмидта столбцов A_s . Как было указано в гл. 4, § 55, на практике это обычно приводит к Q_s , которая отнюдь не ортогональна. В этом случае численная устойчивость, обычно присущая ортогональным преобразованиям подобия, может быть потеряна. На практике обычно определяем такую матрицу Q_s^T , что

$$Q_s^T A_s = R_s. \quad (35.1)$$

Матрица Q_s^T не определяется явно, чаще она определяется в виде произведения либо плоских вращений, либо матриц отражения, используемых при приведении матрицы к треугольному виду по методу Гивенса или Хаусхолдера соответственно. Матрица $R_s Q_s$ вычисляется затем при помощи последовательных умножений R_s справа на транспонированные сомножители для Q_s^T . Мы показали в гл. 3, §§ 26, 35, 45, что произведение вычисленных сомножителей Q_s^T будет почти точно ортогональным и что вычисленная $R_s Q_s$ будет близка к матрице, полученной при точном умножении на вычисленные сомножители. Следовательно, можно показать, что вычисленная A_{s+1} близка к точному ортогональному преобразованию вычисленной A_s . (Мы не можем показать, что вычисленная A_{s+1} близка к такому преобразованию вычисленной A_s , которое получилось бы при точном выполнении s -го шага, и, вообще говоря, это неверно.)

Ясно, что объем работы даже больше, чем при LR-преобразовании. QR-алгоритм имеет практическую ценность, только если A_1 имеет верхнюю форму Хессенберга или симметричную и ленточную форму. Сначала будем иметь дело с матрицами в верхней форме Хессенберга и будем предполагать, что ни один из поддиагональных элементов не равен нулю, так как в противном случае мы можем работать с матрицами Хессенберга более низкого порядка.

Как мы покажем, форма Хессенберга инвариантна по отношению к QR-преобразованию. Рассмотрим сначала приведение к треугольному виду матрицы Хессенберга по методу Гивенса. Легко видеть, что приведение достигается вращениями в плоскостях $(1, 2)$, $(2, 3)$, $\dots$, $(n-1, n)$ соответственно и что требуется только $2n^2$ умножений. Использование метода Хаусхолдера не дает никаких преимуществ, так как на каждом основном шаге приведения обращается в нуль только один элемент. При умножении справа верхней треугольной матрицы на транспонированные матрицы вращений, форма Хессенберга восстанавливается, как и в LR-алгоритме с перестановками. В одном полном шаге QR-алгоритма будет $4n^2$ умножений по сравнению с n^2 в LR с перестановками, но свойства сходимости значительно лучше у QR-алгоритма.

Как и в LR-алгоритме с перестановками, видим, что если мы запоминаем каждую A_s по столбцам, то факторизация и умножение могут быть скомбинированы таким образом, что на каждом шаге A_s считывается из внешней памяти лишь один раз и заменяется на A_{s+1} . Необходимый объем быстродействующей памяти должен быть приблизительно $4n$ слов, для двух столбцов A_s и для косинусов и синусов $(n-1)$ углов вращения.

Сдвиг

36. Так же как и в LR -алгоритме, в QR -алгоритме может быть введен сдвиг с восстановлением или без восстановления. Процесс с восстановлением имеет вид

$$A_s - k_s I = Q_s R_s, \quad R_s Q_s + k_s I = A_{s+1}, \quad (36.1)$$

и, как и раньше, мы можем показать, что

$$A_s = (Q_1 Q_2 \dots Q_{s-1})^T A_1 Q_1 Q_2 \dots Q_{s-1} \quad (36.2)$$

и

$$(Q_1 Q_2 \dots Q_s) (R_s \dots R_2 R_1) = (A_1 - k_1 I) (A_1 - k_2 I) \dots (A_1 - k_s I). \quad (36.3)$$

Эти сдвиги важны на практике. Если A_1 — вещественная и известно, что она имеет вещественные собственные значения, или если A_1 — комплексная, мы можем выбирать сдвиг в точности так же, как в § 24. В первом случае вся арифметика вещественная. В общем случае число итераций при QR -преобразованиях оказалось значительно меньше по сравнению с модифицированным LR -процессом при использовании сдвигов, определенных в каждом случае по одному и тому же принципу.

Если A_1 — вещественная, но имеет несколько комплексно сопряженных собственных значений, то, вообще говоря, на некоторой стадии итераций матрица второго порядка в нижнем правом углу будет иметь комплексно сопряженные собственные значения $\lambda, \bar{\lambda}$. Если мы выполним два шага QR , используя сдвиги λ и $\bar{\lambda}$, то точные вычисления дадут в результате вещественную матрицу, хотя промежуточная матрица будет комплексной. Хотя теперь мы не имеем трудностей, отмеченных в § 27 в связи с перестановками, все же на практике конечная матрица может иметь значительную мнимую часть, хотя собственные значения в точности сохраняются. Френсис описал элегантный метод комбинирования этих двух шагов таким образом, что избегается использование комплексной арифметики. Окончательная матрица получается вещественной, и процесс столь же устойчив, как и стандартный QR -алгоритм. Прежде чем обсуждать этот метод, рассмотрим, однако, два других способа, которые дают значительную экономию в объеме вычислений. Они могут быть использованы как с QR -, так и с модифицированным LR -алгоритмами, и будет проще рассмотреть последний случай.

Разложение матриц A_s

37. Первый способ основан на том, что хотя мы и предполагаем, что A_1 имеет отличные от нуля поддиагональные элементы, мы знаем, что, вообще говоря, некоторые из поддиагональных элементов A_s стремятся к нулю.

Может случиться, что в процессе итераций у матрицы A_s окажется один или более очень малых поддиагональных элементов. Если какой-либо из таких элементов достаточно мал, то мы можем рассматривать его как нуль и иметь дело с меньшими матрицами Хессенберга. Так,

для матрицы

$$A_s = \left[\begin{array}{cc|cccc} \times & \times & & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times \\ \hline & \varepsilon & & \times & \times & \times & \times \\ & & & \times & \times & \times & \times \\ & & & & \times & \times & \times \\ & & & & & \times & \times \\ & & & & & & \times & \times \end{array} \right] \approx \left[\begin{array}{c|c} B & C \\ \hline 0 & D \end{array} \right] \quad (37.1)$$

при достаточно малом ε мы можем найти собственные значения матриц D и B . Заметим, что элементы C в дальнейшем не играют роли. Заметим также, что исходная матрица Хессенберга была сама получена в результате предыдущих вычислений, и следовательно, если ε меньше, чем ошибки в элементах A_1 , то мы мало теряем при замене ε на нуль. Число порядка величины $2^{-t} \|A\|_\infty$ является хорошим допуском.

Итак, на каждой стадии итераций мы можем проверять величины поддиагональных элементов, и если последний из этих элементов, которым можно пренебречь, находится в позиции $(r+1, r)$, то мы можем продолжать итерации с матрицей порядка $n-r$, расположенной в нижнем правом углу, возвращаясь к верхней матрице, когда все собственные значения нижней уже найдены. Если $r = n-1$, нижняя матрица будет первого порядка, и ее собственное значение известно. На этом этапе мы можем выполнить исчерпывание, опустив последние строку в столбец. Если $r = n-2$, нижняя матрица второго порядка и мы можем найти ее собственные значения, решая квадратное уравнение. Снова можно выполнить исчерпывание, опустив последние две строки и два столбца.

38. Второй способ основан на обобщении вышеизложенной идеи Френсиса (1961). Рассмотрим матрицу Хессенберга A , имеющую малые элементы ε_1 и ε_2 в позициях $(r+1, r)$ и $(r+2, r+1)$. Это иллюстрируется в случае $n=6, r=2$ матрицей, имеющей вид

$$A = \left\{ \begin{array}{cc|cccc} \times & \times & & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times \\ \hline & \varepsilon_1 & & \times & \times & \times & \times \\ & & & \varepsilon_2 & \times & \times & \times \\ & & & & \times & \times & \times \\ & & & & & \times & \times \end{array} \right\} = \left[\begin{array}{c|c} X & Y \\ \hline E & W \end{array} \right]. \quad (38.1)$$

Пусть A разделена по отношению к этим элементам способом, показанным в (38.1), где E — нулевая матрица с точностью до элемента ε_1 . Покажем, что любое собственное значение μ матрицы W является также собственным значением матрицы A' , которая отличается от A только элементом $\varepsilon_1 \varepsilon_2 / (a_{r+1, r+1} - \mu)$ в позиции $(r+2, r)$.

Действительно, так как $W - \mu I$ — особенная, найдется отличный от нуля вектор x , для которого

$$x^T (W - \mu I) = 0. \quad (38.2)$$

Первые две компоненты x удовлетворяют соотношению

$$(a_{r+1, r+1} - \mu) x_1 + \varepsilon_2 x_2 = 0, \quad (38.3)$$

которое дает

$$\varepsilon_1 x_1 + [\varepsilon_1 \varepsilon_2 / (a_{r+1, r+1} - \mu)] x_2 = 0. \quad (38.4)$$

Уравнения (38.2) и (38.4) вместе означают, что последние $n - r$ строк модифицированной матрицы $(A' - \mu I)$ линейно зависимы, и следовательно, μ является собственным значением A' . Если $\varepsilon_1 \varepsilon_2 / (a_{r+1, r+1} - \mu)$ достаточно мало, то μ можно считать собственным значением A . Если предположить, что $a_{r+1, r+1} - \mu$ не мало, то малость ε_1 и ε_2 позволяет отделить W от остальной части матрицы.

Численный пример

39. Этот эффект иллюстрируется в табл. 3. Матрица шестого порядка имеет элементы, равные 10^{-5} , в позициях (3, 2) и (4, 3). Собственные значения матрицы четвертого порядка в нижнем правом углу равны $\mu_1, \mu_2, \mu_3, \mu_4$, и только μ_4 близко к a_{33} . Следовательно, мы можем ожидать, что полная матрица имеет собственные значения $\lambda_1, \lambda_2, \lambda_3$, отличающиеся от μ_1, μ_2, μ_3 на величины порядка 10^{-10} . Собственное значение λ_4 отличается от μ_4 на величину порядка 10^{-5} . Собственные значения λ_5 и λ_6 не имеют соответствия, хотя, конечно, они очень близки к собственным значениям угловой главной матрицы третьего порядка.

		Таблица 3					
$A_1 = \left[\begin{array}{c c} X & Y \\ \hline E & W \end{array} \right] =$	4	3	2	1	5	6	
	3	1	4	2	1	7	
	10^{-5}		6	1	5	3	
			10^{-5}	2	5	7	
				1	2	3	
					4	1	
Собственные значения A_1		Собственные значения W					
$\lambda_1 = -0,46410\ 31621$		$\mu_1 = -0,46410\ 31621$					
$\lambda_2 = -0,99999\ 85712$		$\mu_2 = -0,99999\ 85714$					
$\lambda_3 = 6,46412\ 31631$		$\mu_3 = 6,46412\ 31611$					
$\lambda_4 = 6,00011\ 84534$		$\mu_4 = 5,99997\ 85724$					
$\lambda_5 = -0,85410\ 48844$							
$\lambda_6 = 5,85396\ 50012$							

Практическая процедура

40. Для использования преимуществ, вытекающих из этого явления при применении LR-алгоритма, мы поступали следующим образом. В конце каждой итерации мы определяли сдвиг k_s , который будет использован на следующем этапе. Затем проверяли, удовлетворяется ли условие

$$a_{r+1, r}^{(s)} a_{r+2, r+1}^{(s)} / (a_{r+1, r+1}^{(s)} - k_s) < 2^{-t} \|A_1\| \quad (40.1)$$

при $r = 1, 2, \dots, n - 1$, и обозначали наибольшее значение r , для которого это верно, через p . Затем выполняли разложение в произведение треугольных с перестановками матрицы $(A_s - k_s I)$, начиная с $(p + 1)$ -й строки, а не с первой. При этом в позиции $(p + 2, p)$ оказывается элемент $a_{p+1, p}^{(s)} a_{p+2, p+1}^{(s)} / (a_{p+1, p+1}^{(s)} - k_s)$; на первом шаге не понадобится перестановки, и при наших предположениях этим элементом можно пренебречь. В конце частичной факторизации матрица будет иметь вид для

случая $n = 6, p = 2$

$$\left[\begin{array}{cc|cccc} \times & \times & & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times \\ \hline & \varepsilon_1 & & \times & \times & \times & \times \\ & - & & 0 & \times & \times & \times \\ & & & & 0 & \times & \times \\ & & & & & 0 & \times \end{array} \right]. \quad (40.2)$$

Здесь первые p строк остались неизменными. Черточка означает пренебрежимо малый элемент, введенный при приведении. Теперь мы завершаем преобразование подобия. Очевидно, что при этом будут производиться действия только со столбцами от $p + 1$ до n , включая элементы матрицы в верхнем правом углу. Обычно, если условие (40.1) удовлетворялось на s -м шаге, то оно выполнялось и на последующих шагах.

Мы описали процесс для LR -алгоритма с перестановками, но все немедленно переносится и на QR -алгоритм. На самом деле достаточно того же самого условия, так как если оно удовлетворяется, мы можем начинать приведение к треугольному виду с $(p + 1)$ -й строки. Угол первого вращения мал, и в позиции $(p + 2, p)$ получается пренебрежимо малый элемент $\varepsilon_1 \varepsilon_2 / |a_{p+1, p+1}^{(s)} - k_s|^2 + \varepsilon_2^2|^{1/2}$.

Очевидно, можно комбинировать оба описанных способа, используя преимущества малости как $a_{p+1, p}^{(s)}$, так и $a_{p+1, p}^{(s)} a_{p+2, p+1}^{(s)} / (a_{p+1, p+1}^{(s)} - k_s)$.

Устранение комплексно сопряженных сдвигов

41. Рассмотрим теперь случай вещественных матриц, имеющих несколько комплексно сопряженных собственных значений. Будем пока рассматривать только QR -алгоритм. Предположим, что на s -м шаге собственные значения матрицы второго порядка в нижнем правом углу равны k_1 и k_2 . Они могут быть вещественными или комплексно сопряженной парой. В обоих случаях рассмотрим эффект от выполнения двух шагов QR со сдвигами соответственно k_1 и k_2 . Для удобства будем обозначать s -ю матрицу через A_1 . Имеем

$$A_1 - k_1 I = Q_1 R_1, \quad R_1 Q_1 + k_1 I = A_2, \quad (41.1)$$

$$A_2 - k_2 I = Q_2 R_2, \quad R_2 Q_2 + k_2 I = A_3, \quad (41.2)$$

$$Q_1 Q_2 (R_2 R_1) = (A_1 - k_1 I) (A_1 - k_2 I), \quad (41.3)$$

$$A_3 = (Q_1 Q_2)^H A_1 Q_1 Q_2. \quad (41.4)$$

Если k_1 и k_2 не являются точными собственными значениями, QR -разложения единственны, если мы потребуем, чтобы верхние треугольные матрицы имели положительные диагональные элементы. Матрица справа в (41.3) всегда вещественная, и следовательно, $Q_1 Q_2$ — вещественная матрица, полученная при ортогональном приведении матрицы $(A_1 - k_1 I) (A_1 - k_2 I)$ к треугольному виду. Положим

$$Q_1 Q_2 = Q, \quad R_2 R_1 = R, \quad QR = (A_1 - k_1 I) (A_1 - k_2 I). \quad (41.5)$$

Уравнение (41.4) означает, что Q — такая вещественная матрица, что

$$A_1 Q = Q A_3. \quad (41.6)$$

Предположим теперь, что каким-либо способом мы можем определить ортогональную матрицу $\tilde{Q}$, первый столбец которой совпадает с первым столбцом Q и

$$A_1 \tilde{Q} = \tilde{Q} B, \quad (41.7)$$

где B — матрица Хессенберга с положительными поддиагональными элементами. Из § 7 гл. 6 знаем, что такая $\tilde{Q}$ должна совпадать с Q , а B с A_3 . Мы хотим определить эти $\tilde{Q}$ и B без использования комплексной арифметики. В самом деле, будем делать это, вычисляя $\tilde{Q}^T$, которая имеет ту же первую строку, что и Q^T , и такую, что

$$\tilde{Q}^T A_1 = B \tilde{Q}^T. \quad (41.8)$$

Из (41.5) имеем

$$Q^T (A_1 - k_1 I) (A_1 - k_2 I) = R, \quad (41.9)$$

и, следовательно, Q^T — ортогональная матрица, которая приводит $(A_1 - k_1 I) (A_1 - k_2 I)$ к верхней треугольной матрице. По методу Гивенса для приведения к треугольному виду Q^T определяется в виде произведения $\frac{1}{2} n (n - 1)$ плоских вращений R_{ij} . Следовательно,

$$Q^T = R_{n-1, n} \dots R_{2, n} \dots R_{2, 3} R_{1, n} \dots R_{1, 3} R_{1, 2}. \quad (41.10)$$

Если будем строить произведение, начиная с $R_{1, 2}$, мы найдем, что первая строка Q^T равна первой строке $R_{1, n} \dots R_{1, 3} R_{1, 2}$, и последующие умножения не меняют эту строку. Вращения $R_{1, 2}, R_{1, 3}, \dots, R_{1, n}$ полностью определяют первый столбец $(A_1 - k_1 I) (A_1 - k_2 I)$ и, следовательно, могут быть вычислены, даже если остальная часть матрицы неизвестна. Первый столбец имеет лишь три отличных от нуля элемента. Обозначим их x_1, y_1, z_1 . Тогда

$$\left. \begin{aligned} x_1 &= (a_{11} - k_1) (a_{11} - k_2) + a_{12} a_{21} = \\ &= a_{11}^2 + a_{12} a_{21} - a_{11} (k_1 + k_2) + k_1 k_2, \\ y_1 &= a_{21} (a_{11} - k_2) + (a_{22} - k_1) a_{21} = a_{21} (a_{11} + a_{22} - k_1 - k_2), \\ z_1 &= a_{32} a_{21}. \end{aligned} \right\} \quad (41.11)$$

Поэтому $R_{1, 4}, R_{1, 5}, \dots, R_{1, n}$ суть единичные матрицы, а $R_{1, 2}$ и $R_{1, 3}$ определяются по x_1, y_1, z_1 и, очевидно, вещественные. Предположим, что $R_{1, 2}$ и $R_{1, 3}$ вычислены, и определим матрицу

$$C_1 = R_{1, 3} R_{1, 2} A_1 R_{1, 2}^T R_{1, 3}^T. \quad (41.12)$$

В гл. 6, § 7 мы показали, что для любой вещественной матрицы C_1 существует такая ортогональная матрица S_1 , первый столбец которой равен e_1 , что

$$S_1^T C_1 S_1 = B, \quad (41.13)$$

где B — верхняя матрица Хессенберга. Матрица S_1 может быть определена при помощи процессов Гивенса или Хаусхолдера. Следовательно,

$$(S_1^T R_{1, 3} R_{1, 2}) A_1 (R_{1, 2}^T R_{1, 3}^T S_1) = B, \quad (41.14)$$

и если мы обозначим $(S_1^T R_{1, 3} R_{1, 2}) = \tilde{Q}^T$, то из (41.14) имеем

$$\tilde{Q}^T A_1 = B \tilde{Q}^T. \quad (41.15)$$

Так как первая строка S_1^T равна e_1^T , то первая строка $\tilde{Q}^T$ совпадает с первой строкой $R_{1, 3} R_{1, 2}$, которая, как было показано, совпадает с первой

строкой Q^T . Следовательно, B должна быть равна A_3 , которая получилась бы при выполнении двух шагов QR со сдвигами k_1 и k_2 . Мы получили метод определения A_3 без использования комплексной арифметики даже в случае, когда k_1 и k_2 — комплексные.

42. Как мы сейчас покажем, этот способ очень эффективен. Число умножений при вычислении вращений $R_{1,2}$ и $R_{1,3}$ не зависит от n . После того как они получены, по (41.12) вычисляется C_1 , и снова число умножений не зависит от n . В случае $n = 8$ C_1 имеет вид

$$C_1 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ \times & 0 & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \end{bmatrix},$$

$$C_3 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & \times & 0 & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \end{bmatrix} \quad \begin{matrix} \leftarrow \\ \leftarrow \\ \leftarrow \end{matrix} \quad (42.1)$$

↑ ↑ ↑

Аннулирование элемента, обозначенного нулем, не дает никаких преимуществ. Матрица C_1 уже не будет в форме Хессенберга, так как при умножении слева и справа появились в столбцах 1 и 2 отличные от нуля элементы. Рассмотрим теперь подобное преобразование C_1 к форме Хессенберга методом Гивенса. Так как C_1 почти в форме Хессенберга, то объем работы значительно меньше, чем если бы C_1 была полной матрицей.

Как и в общем случае (гл. 6, § 2), здесь всего $n - 2$ основных шага, причем на r -м шаге аннулируются элементы в r -м столбце. Мы сейчас видим, что перед началом r -го шага C_r имеет форму Хессенберга, за исключением ненулевых элементов в позициях $(r + 2, r)$ и $(r + 3, r)$. Элемент $(r + 3, r + 1)$ равен нулю, но это не дает никаких преимуществ. Это иллюстрируется в случае $n = 8$, $r = 3$ в (42.1). Предполагая, что на r -м шаге мы имеем такую конфигурацию, умножаем слева на вращения в плоскостях $(r + 1, r + 2)$ и $(r + 1, r + 3)$ так, чтобы аннулировать два подчеркнутых элемента в (42.1). Умножение слева уничтожит нуль в позиции $(r + 3, r + 1)$, но больше не будет введено никаких новых ненулевых элементов. Затем умножаем справа на транспонированные матрицы вращений. В результате столбцы $r + 1$ и $r + 2$ заменятся их линейными комбинациями, аналогично столбцы $r + 1$ и $r + 3$ — их линейными комбинациями. В (42.1) мы отметили стрелками строки и столбцы C_r , которые изменяются на r -м шаге. Очевидно, что форма C_{r+1} точно аналогична форме C_r .

Мы описали предварительное преобразование A_1 к C_1 так, как будто оно отлично от преобразования, которое требуется для приведения C_1 .

Но, если мы припишем к A_1 еще один столбец, состоящий из элементов x_1, y_1, z_1 , как показано в (42.2)

$$C_0 = \begin{bmatrix} x_1 & \times & \times & \times & \times & \times & \times & \times & \times \\ y_1 & \times & \times & \times & \times & \times & \times & \times & \times \\ z_1 & 0 & \times & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times & \times \\ & & & & & \times & \times & \times & \times \\ & & & & & & \times & \times & \times \\ & & & & & & & \times & \times \end{bmatrix} = \begin{bmatrix} x_1 & \vdots \\ y_1 & \vdots \\ z_1 & \vdots \\ 0 & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & \vdots \end{bmatrix} A_1, \quad (42.2)$$

и обозначим эту расширенную матрицу C_0 , то C_1 получается из C_0 при помощи шагов, которые в точности аналогичны шагам, необходимым для получения C_{r+1} по C_r . Для перехода от A_1 к A_3 требуется $8n^2$ умножений, столько же, сколько для двух вещественных шагов QR-алгоритма. Если бы мы осуществляли два шага с комплексными сдвигами k_1 и k_2 с использованием комплексной арифметики, потребовалось бы $16n^2$ умножений.

Двойной QR-шаг с использованием матриц отражения

43. Мы описали двойной QR-шаг в терминах преобразования Гивенса. В случае одного сдвига метод Хаусхолдера не лучше, но в случае двойного сдвига он значительно более экономен.

Мы используем обозначение $P_r = I - 2w_r w_r^T$, где

$$w_r^T = (0, \underbrace{0, \dots, 0}_{r-1}, \times, \times, \dots, \times), \quad (43.1)$$

и покажем, что A_3 может быть получена следующим способом.

Сначала находится такая матрица отражения P_1 , что

$$P_1 x = k e_1, \quad \text{где} \quad x^T = (x_1, y_1, z_1, 0, \dots, 0). \quad (43.2)$$

Очевидно, что только три компоненты соответствующего w_1 отличны от нуля. Затем вычисляется матрица $C_1 = P_1 A_1 P_1$, которая имеет форму

$$C_1 = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \end{bmatrix},$$

$$C_1^T = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \end{bmatrix}. \quad (43.3)$$

Эта форма почти совпадает с формой матрицы C_1 , полученной при помощи плоских вращений (см. (42.1)). Единственное отличие состоит в том, что теперь уже нет нуля в позиции (4.2), но это не имеет особого значения.

Матрица C_1 затем приводится к форме Хессенберга методом Хаусхолдера (гл. 6, § 3). Это производится при помощи последовательных умножений слева и справа на матрицы $P_2, P_3, \dots, P_{n-2}$.

Резльтирующая матрица B Хессенберга такова, что

$$P_{n-2} \dots P_2 P_1 A_1 P_1 P_2 \dots P_{n-2} = B \quad (43.4)$$

и $P_{n-2} \dots P_1$ имеет ту же первую строку, что и P_1 , первая строка которой в свою очередь совпадает с первой строкой ортогональной матрицы, которая приводит

$$(A - k_1 I)(A - k_2 I)$$

к треугольному виду. Следовательно, B равна A_3 .

Покажем, что перед умножением слева на P_r текущая матрица C_{r-1} имеет форму, показанную в (43.3) для случая $n = 8, r = 4$. Действительно, предположим, что это верно для C_{r-1} . При помощи P_r аннулируются только два элемента в столбце $r - 1$, и следовательно, соответствующий w_r имеет отличные от нуля элементы только в позициях $r, r + 1, r + 2$. Умножение справа и слева на P_r заменяет строки и столбцы $r, r + 1, r + 2$ их линейными комбинациями, и таким образом аннулируются элементы в позициях $(r + 1, r - 1)$ и $(r + 2, r - 1)$. Следовательно, матрица C_r имеет ту же форму, что и C_{r-1} , и так как C_1 нужного вида, результат установлен. Как и в § 42, мы можем включить начальное преобразование в общую схему при помощи добавления к матрице A_1 дополнительного первого столбца, состоящего из элементов x_1, y_1, z_1 .

Вычислительные детали

44. Ненулевые элементы P_r определяются элементами $c_{r, r-1}^{(r-1)}, c_{r+1, r-1}^{(r-1)}, c_{r+2, r-1}^{(r-1)}$. Будем обозначать эти элементы x_r, y_r, z_r , что соответствует нашему предыдущему обозначению x_1, y_1, z_1 . Наиболее экономичное представление P_r описано в гл. 5, § 33. Оно дает

$$\left. \begin{aligned} P_r &= I - 2p_r p_r^T / \|p_r\|_2^2, \\ p_r^T &= (0, \dots, 0, 1, u_r, v_r, 0, \dots, 0), \\ S_r^2 &= x_r^2 + y_r^2 + z_r^2, \quad u_r = y_r / (x_r \mp S_r), \quad v_r = z_r / (x_r \mp S_r), \\ 2/\|p_r\|_2^2 &= 2/(1 + u_r^2 + v_r^2) = \alpha_r, \end{aligned} \right\} \quad (44.1)$$

где знак выбран так, что

$$|x_r \mp S_r| = |x_r| + S_r. \quad (44.2)$$

При таком методе вычислений P_r ортогональна с рабочей точностью. Используем P_r в виде

$$P_r = I - p_r (\alpha_r p_r^T). \quad (44.3)$$

Это дает то преимущество, что одна из компонент p_r равна единице. При таком способе перехода от A_1 к A_3 требуется только $5n^2$ умножений.

Разложение матриц A_s .

45. При использовании двухшаговой техники мы также можем использовать способы, описанные в §§ 37, 38. Если поддиагональный элемент на некоторой стадии с рабочей точностью равен нулю, то это очевидно. В случае, когда два последовательных поддиагональных элемента A_s «малы», модификация не столь тривиальна. Если $a_{r+1, r}^{(s)}$ и $a_{r+2, r+1}^{(s)}$ достаточно малы, мы надеемся начинать преобразования с P_{r+1} вместо P_1 . По аналогии с (41.11) вычисляем x_{r+1} , y_{r+1} , z_{r+1} вида

$$\left. \begin{aligned} x_{r+1} &= a_{r+1, r+1}^2 + a_{r+1, r+2} a_{r+2, r+1} - a_{r+1, r+1} (k_1 + k_2) + k_1 k_2, \\ y_{r+1} &= a_{r+2, r+1} (a_{r+1, r+1} + a_{r+2, r+2} - k_1 - k_2), \\ z_{r+1} &= a_{r+3, r+2} a_{r+2, r+1} \end{aligned} \right\} \quad (45.1)$$

и по ним определяем P_{r+1} обычным способом. В случае $n = 7$, $r = 2$ матрицы P_{r+1} и A_s имеют вид

$$P_3 = \begin{bmatrix} I & & \\ & I - 2w_3 w_3^T & \end{bmatrix},$$

$$A_s = \begin{bmatrix} \times & \times & & \times & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times & \times \\ \varepsilon_1 & & \times & \times & \times & \times & \times & \times \\ & \varepsilon_2 & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \end{bmatrix} = \begin{bmatrix} X & Y \\ E & Z \end{bmatrix}, \quad (45.2)$$

где мы обозначили $a_{r+1, r}$ и $a_{r+2, r+1}$ через ε_1 и ε_2 для того, чтобы подчеркнуть, что они малы. Умножение слева на P_{r+1} не меняет первые r строк. Часть, обозначенная E в (45.2), имеет только один отличный от нуля элемент и, вообще говоря, после умножения слева будет иметь три ненулевых элемента, как показано в (45.3):

$$\begin{bmatrix} \times & \times & & \times & \times & \times & \times & \times \\ \times & \times & & \times & \times & \times & \times & \times \\ \eta_1 & & \times & \times & \times & \times & \times & \times \\ \eta_2 & & \times & \times & \times & \times & \times & \times \\ \eta_3 & & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \end{bmatrix}; \quad (45.3)$$

у матрицы Z изменятся строки $r+1$, $r+2$, $r+3$ обычным образом. Мы требуем, чтобы η_2 и η_3 были пренебрежимо малыми. В обозначениях § 44 имеем

$$\left. \begin{aligned} \eta_1 &= \varepsilon_1 - 2\varepsilon_1/(1 + u_{r+1}^2 + v_{r+1}^2), \\ \eta_2 &= -2\varepsilon u_{r+1}/(1 + u_{r+1}^2 + v_{r+1}^2), \\ \eta_3 &= -2\varepsilon v_{r+1}/(1 + u_{r+1}^2 + v_{r+1}^2). \end{aligned} \right\} \quad (45.4)$$

Из (45.4) мы видим, что y_{r+1} и z_{r+1} имеют множитель $a_{r+2, r+1} = \varepsilon_2$, а x_{r+1} не будет, вообще говоря, малым. Следовательно, u_{r+1} и v_{r+1} имеют

множитель ε_2 , так что η_2 и η_3 имеют множитель $\varepsilon_1 \varepsilon_2$, а η_1 будет равен $-\varepsilon_1$, если ε_2 достаточно мало. Нет надобности вычислять η_2 и η_3 точно. Достаточным условием того, что можно начинать с $(r+1)$ -й строки, является следующее:

$$|\varepsilon_1(|y_{r+1}| + |z_{r+1}|)/x_{r+1}| < 2^{-t} \|A_1\|_E. \quad (45.5)$$

Заметим, что для того чтобы получить ε_1 , надо изменить знак η_1 .

Обычно очень важно при обоих методах, которые мы описали, полностью использовать преимущества малости поддиагональных элементов. К сожалению, надо отметить, что при применении этих приемов нельзя думать только о сокращении количества умножений. Приведение матрицы к форме Хессенберга единственно, только если поддиагональные элементы отличны от нуля, и поэтому важно использовать только нижнюю подматрицу, если поддиагональный элемент пренебрежимо мал. В этом случае верхняя и нижняя половины матрицы слабо связаны. Если некоторые собственные значения верхней половины меньше собственных значений нижней половины, то при достаточно большом количестве итераций с нулевым сдвигом малое собственное значение верхней половины будет проникать в нижний правый угол, но из-за того, что связь слабая, этот процесс будет очень медленным. Применение способов, которые мы описали, помогает избежать потери времени, но при этом собственные значения не получаются в порядке возрастания их абсолютной величины.

Прием двойного сдвига для LR

46. Можно описать, пренебрегая перестановками, подобный прием двойного сдвига для LR -алгоритма. Аналогичные (41.1) и (41.4) уравнения будут

$$L_1 L_2 R_2 R_1 = (A_1 - k_1 I)(A_1 - k_2 I), \quad A_1 L_1 L_2 = L_1 L_2 A_3 \quad (46.1)$$

и, обозначив $L_1 L_2 = L$, $R_2 R_1 = R$, имеем

$$LR = (A_1 - k_1 I)(A_2 - k_2 I), \quad A_1 \tilde{L} = L A_3. \quad (46.2)$$

Теперь используется тот факт, что если $\tilde{L}$ — единичная нижняя треугольная матрица, первый столбец которой равен первому столбцу L и

$$A_1 \tilde{L} = \tilde{L} B, \quad (46.3)$$

где B — верхняя матрица Хессенберга, то $\tilde{L}$ равна L , а B равна A_3 (гл. 6, § 11). По аналогии с алгоритмом QR будем вместо сомножителей $\tilde{L}$ искать сомножители $\tilde{L}^{-1}$.

Первый столбец $(A_1 - k_1 I)(A_1 - k_2 I)$, так же как и раньше, имеет только три отличных от нуля элемента x_1, y_1, z_1 , выписанные в (41.11). Определим далее элементарную матрицу типа N_1 , которая приводит первый столбец к кратному e_1 . Очевидно, что первый столбец N_1^{-1} будет первым столбцом L . Наконец, вычисляем C_1 , равную $C_1 = N_1 A_1 N_1^{-1}$, и приводим ее к форме Хессенберга, используя элементарные неунитарные матрицы типа $N_2, N_3, \dots, N_{n-2}$ (гл. 6, § 8). Следовательно,

$$N_{n-2} \dots N_2 N_1 A_1 N_1^{-1} N_2^{-1} \dots N_{n-2}^{-1} = B, \quad (46.4)$$

и матрица $N_1^{-1} N_2^{-1} \dots N_{n-2}^{-1}$ это единичная нижняя треугольная матрица, первый столбец которой совпадает с первым столбцом N_1^{-1} и, следовательно,

с первым столбцом L . Поэтому B равна A_3 . На r -м этапе приведения текущая матрица имеет в точности ту же форму, что и матрица при использовании приведения Гивенса на соответствующей стадии, показанной в (42.1). Следовательно, каждая N_r , за исключением последней, имеет только два ненулевых элемента под диагональю, которые находятся в позициях $(r+1, r)$ и $(r+2, r)$. Так же, как и в QR , начальный шаг может быть скомбинирован с остальными. Полное число умножений порядка $2n^2$.

На всех этапах мы можем применять перестановки, заменяя N_r , и N_1 , на $N_r I_r$, r' , но теоретическое обоснование процесса при этом несколько затруднительно. Перестановки в первых строках и столбцах продолжаются, вообще говоря, даже тогда, когда некоторые из остальных собственных значений почти определены. Тем не менее мы находим, что способ двойного сдвига работает весьма удовлетворительно, хотя абсолютно необходимо, чтобы пренебрежимо малые поддиагональные элементы и малые последовательные поддиагональные элементы определялись так, как было описано.

Сравнение LR- и QR-алгоритмов

47. Сейчас сравним относительные достоинства LR- и QR-алгоритмов при приложении их к матрицам Хессенберга. По моему мнению, следует отдать предпочтение унитарным преобразованиям со значительно большей уверенностью, чем в предыдущих случаях, когда мы сталкивались с подобным выбором. Свойства сходимости LR-алгоритма с перестановками далеко не удовлетворительны, и требуется довольно значительное число преобразований, так что обусловленность матриц может ухудшаться, а недиагональные элементы увеличиваться. С другой стороны, для QR-алгоритма легко показать, используя метод гл. 3, § 45, что A_s в точности унитарно подобна $(A_1 + E_s)$, где E_s мала. Например, можно показать, что при использовании арифметики с плавающей запятой с накоплением скалярных произведений, где это нужно, $\|E_s\|_E < Ksn2^{-t} \|A_1\|_E$. Более того, оказалось, что количество итераций, необходимых при использовании QR, как правило меньше, чем при LR.

Оказалось, что процедура, основанная на приеме двойного сдвига для QR, со сдвигами, определенными из нижней матрицы второго порядка, и включающая определение малых поддиагональных элементов, является наиболее мощной программой широкого назначения из всех, которые я использовал. Ввиду важности этого процесса стоит коснуться некоторых неправильных представлений о его сходимости.

В § 32 мы заметили, что если матрица A_1 имеет r собственных значений, равных по абсолютной величине, то на каждом этапе итераций при отсутствии сдвига на диагонали могут возникнуть квадратные блоки порядка r . Это может создать впечатление, что отделение таких групп представляет трудности. Однако при использовании сдвигов это обычно неверно. Например, для каждой из матриц

$$\begin{array}{cc} \text{(i)} & \text{(ii)} \\ \left[\begin{array}{ccccc} -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \end{array} \right], & \left[\begin{array}{ccccc} 0 & -1 & -1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \end{array} \right], \end{array}$$

$$\begin{aligned}
 & \text{(iii)} \quad \begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & -1 \\ & & 1 & 0 & -1 \\ & & & 1 & 0 \end{bmatrix}, \quad \text{(iv)} \quad \begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -\frac{1}{2} \\ & 1 & 0 & 0 & -1 \\ & & 1 & 0 & -1 \\ & & & 1 & -\frac{1}{2} \end{bmatrix}, \\
 & \text{(v)} \quad \begin{bmatrix} -\frac{1}{2} & -1 & -1 & -\frac{1}{2} & -1 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix} \quad (47.1)
 \end{aligned}$$

был применен обычный способ двойного сдвига для QR . Хотя модуль собственных значений матриц (i), (ii) и (iii) равен единице, среднее число итераций на собственное значение меньше трех. Сдвиги определяются стандартным способом, и модифицированные матрицы уже не имеют равных по модулю собственных значений, так что все трудности пропадают. Это не так для матрицы (32.6), характеристическое уравнение которой есть $\lambda^4 + 1 = 0$. Эта матрица инвариантна по отношению к QR , и оба сдвига, определенные обычным способом, равны нулю. Однако если мы выполним одну итерацию с произвольным сдвигом, то процесс будет сходиться. Кроме того, если $(\lambda^r + 1)$ — сомножитель в характеристическом уравнении, то это еще не означает, что нули этого сомножителя не будут разделяться при помощи QR -алгоритма. Если предположить, что другие сомножители таковы, что применяются ненулевые сдвиги, то процесс будет сходиться. Пример такого случая дает вторая матрица. Характеристический полином равен $(\lambda^3 + 1)(\lambda^2 + 1)$, но после первой итерации обычный процесс определяет ненулевой сдвиг. Все пять собственных значений были определены с 12 десятичными знаками за 12 итераций.

Кратные собственные значения

48. Вернемся к случаю кратных собственных значений. Соображения § 32 показывают, что кратные собственные значения, которым соответствуют линейные элементарные делители, не влияют неблагоприятно на скорость сходимости. Это иногда вызывает недоумение и дает основания для аргументов следующего типа.

Пусть матрица A_1 имеет элементарные делители $(\lambda - a)$, $(\lambda - a)$ и $(\lambda - b)$, где a и b — приблизительно 1 и 3, и предположим, что на некотором этапе

$$A_r = \begin{bmatrix} 3 & 1 & 2 \\ \varepsilon_1 & 1 & 2 \\ \varepsilon_2 & \varepsilon_3 & 1 \end{bmatrix}. \quad (48.1)$$

Можно проверить, что один шаг LR даст

$$A_{r+1} = \begin{bmatrix} 3 & 1 & 2 \\ \frac{1}{3}\varepsilon_1 + \frac{2}{3}\varepsilon_2 & 1 & 2 \\ \frac{1}{3}\varepsilon_3 & \varepsilon_3 + \frac{1}{3}\varepsilon_2 & 1 \end{bmatrix} \quad (48.2)$$

(QR -алгоритм даст в основном то же самое). Очевидно, что элемент (3, 2) сходится неудовлетворительно.

Ошибка состоит в том, что если A_1 имеет линейные элементарные делители, то мы не можем получить таких A_r . Ранг $(A - aI)$ должен быть равен единице, и следовательно, например, элемент (2, 3) должен удовлетворять соотношению $(3 - a)a_{23} - 2\varepsilon_1 = 0$, так что a_{23} должен быть приблизительно равным ε_1 . Если учесть соотношения, которые должны быть между элементами A_r , то видно, что каждый поддиагональный элемент уменьшается на каждом шаге приблизительно в три раза. Использование сдвигов значительно увеличивает скорость сходимости (ср. § 54 для симметричных матриц).

Если A_1 имеет нелинейные элементарные делители, то LR -алгоритм, вообще говоря, не дает сходимости к матрице в верхней треугольной форме. Так, например, матрица

$$A_1 = \begin{bmatrix} a & 0 & 0 \\ 1 & a & 0 \\ 0 & 1 & a \end{bmatrix} \quad (48.3)$$

инвариантна по отношению к обычной LR -технике, и когда $|a| > 1$, инвариантность имеет место, даже если использовать перестановки. (Тот факт, что A_1 — нижняя треугольная, несуществен для процесса, приводящего к появлению верхней треугольной формы.)

Тем не менее QR -алгоритм сходится. На самом деле мы можем доказать сходимость для любых матриц, имеющих жорданову форму A_1 из (48.3). Действительно, имеем

$$A_1 = X \begin{bmatrix} a & 0 & 0 \\ 1 & a & 0 \\ 0 & 1 & a \end{bmatrix} Y = XJY, \quad A_1^s = X \begin{bmatrix} a^s & 0 & 0 \\ \begin{pmatrix} s \\ 1 \end{pmatrix} a^{s-1} & a^s & 0 \\ \begin{pmatrix} s \\ 2 \end{pmatrix} a^{s-2} & \begin{pmatrix} s \\ 1 \end{pmatrix} a^{s-1} & a^s \end{bmatrix} Y, \quad (48.4)$$

и, следовательно, если $Y = LU$, то имеем

$$\begin{aligned} A_1^s &= X \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \times \\ &\times \begin{bmatrix} \begin{pmatrix} s \\ 2 \end{pmatrix} a^{s-2} + \begin{pmatrix} s \\ 1 \end{pmatrix} l_{21} a^{s-1} + l_{31} a^s & \begin{pmatrix} s \\ 1 \end{pmatrix} a^{s-1} + l_{32} a^s & a^s \\ \begin{pmatrix} s \\ 1 \end{pmatrix} a^{s-1} + l_{21} a^s & a^s & 0 \\ a^s & 0 & 0 \end{bmatrix} U = \\ &= XPK_s U. \end{aligned} \quad (48.5)$$

Если мы теперь положим

$$K_s = L_s U_s, \quad (48.6)$$

то из формы K_s очевидно, что $L_s \rightarrow I$. Следовательно,

$$A_1^s = X P L_s U_s U = X P (I + E_s) U_s U, \quad E_s \rightarrow 0, \quad (48.7)$$

что показывает, как обычно, что в QR -алгоритме последовательность произведений $Q_1 Q_2 \dots, Q_s$ стремится в основном к матрице, полученной при QR -факторизации XP . Если

$$XP = QR, \quad (48.8)$$

то

$$A_s \rightarrow Q^T A_1 Q = Q^T X J X^{-1} Q = R P^T J P R^{-1}, \quad (48.9)$$

и поскольку

$$P^T J P = \begin{bmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{bmatrix}, \quad (48.10)$$

то A_s стремится к верхней треугольной матрице.

Очевидно, что это доказательство можно распространить на случай элементарных делителей любой степени, и затем, как в §§ 30—32, перенести на случай, когда A_1 имеет сложную жорданову форму.

Скорость сходимости определяется скоростью, с которой a/s стремится к нулю, и очевидно, что если a мало, то предел достигается более быстро. Если a достаточно мало, то для матрицы (48.3) достаточно двух итераций. Можно ожидать, что на практике использование сдвигов приводит к тому, что требуется удивительно мало итераций. В качестве примера возьмем матрицу

$$A_1 = \begin{bmatrix} -5 & -9 & -7 & -2 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 & 0 \end{bmatrix}, \quad \det(A_1 - \lambda I) = (\lambda + 1)^3 (\lambda + 2), \quad (48.11)$$

которая имеет кубический элементарный делитель. При работе с 12 десятичными знаками, при использовании QR -алгоритма с двойным сдвигом, требуется только одиннадцать итераций. Конечно, эта матрица плохо обусловлена. Изменение порядка ϵ в элементах A_1 вызывает изменения порядка $\epsilon^{1/3}$ в собственных значениях. Кратные собственные значения имеют погрешность порядка 10^{-4} , а простое собственное значение порядка 10^{-12} . Мы должны рассматривать этот результат как «наилучший возможный» при данной точности вычислений. Для больших матриц с нелинейными элементарными делителями способ двойного сдвига может значительно медленнее давать правильное значение сдвига, и сходимость может быть с самого начала очень медленной.

Строго говоря, наши верхние матрицы Хессенберга никогда не будут иметь линейных элементарных делителей, соответствующих кратным собственным значениям, так как мы предположили, что все поддиагональные элементы не нули, и следовательно, ранг $(A - \lambda I)$ равен $n - 1$ для любого значения λ . Собственному значению кратности r поэтому должен соответствовать делитель степени r . Однако, как мы показали в гл. 5, § 45, существуют матрицы, имеющие патологически близкие собственные

значения, поддиагональные элементы которых не малы. Если предположить, что матрица не такова, что малые возмущения могут привести к нелинейным элементарным делителям, поведение в смысле сходимости будет в основном такое же, как в случае кратных корней, которым соответствуют линейные элементарные делители.

Из нашего обсуждения следует, что, как можно было ожидать, привлечение сдвигов делает QR исключительно эффективным почти для всех матриц. Мой опыт говорит, что это действительно так.

Специальное использование процесса исчерпывания

49. Сейчас в первую очередь будем рассматривать QR -процесс как метод исчерпывания. Если λ_1 — точное собственное значение, то один шаг QR -процесса, примененного к $A_1 - \lambda_1 I$, приводит к матрице, последняя строка которой будет нулевой, и мы можем провести исчерпывание. Предположим теперь, что мы нашли значение λ_1 , верное с рабочей точностью. Можем ли мы выполнить один шаг QR и провести исчерпывание?

Ответом на этот вопрос будет резкое «нет». В самом деле, мы построили очень хорошо обусловленные матрицы, для которых элемент (n, n) преобразованной матрицы без восстановления сдвига, все еще далекий от нуля, является даже наибольшим элементом всей матрицы! Примером такой матрицы является матрица W_{21} . Мы уже обсуждали приведение $W_{21} - (\lambda_1 + \varepsilon) I$ к треугольному виду в гл. 5, § 56 и видели, что последний элемент больше 21, когда ε порядка 10^{-7} .

На первый взгляд это весьма тревожно и может посеять сомнения в численной устойчивости процесса. Однако такие сомнения не обоснованы. Вычисленная преобразованная матрица всегда точно подобна матрице, очень близкой к исходной и, следовательно, собственные значения сохраняются. Если мы продолжим итерации, используя тот же сдвиг, мы сможем в конечном счете произвести исчерпывание. На практике примеры матриц, для которых недостаточно двух итераций, весьма редки. Например, две итерации с матрицей W_{21} дают матрицу, у которой элементы в позициях $(21, 20)$ и $(21, 21)$ равны нулю с рабочей точностью. Мы можем противопоставить это исчерпыванию, изученному в гл. 7, § 53, где неудачное аннулирование величин, которые должны быть точно нулями, было роковым.

Симметричные матрицы

50. Другим классом матриц, к которым применимы LR - и QR -алгоритмы, являются симметричные ленточные матрицы. Рассмотрим сначала общие симметричные матрицы. Очевидно, что QR -алгоритм сохраняет симметричность, так как

$$A_2 = Q_1^T A_1 Q_1, \quad (50.1)$$

но это, вообще говоря, неверно для LR -алгоритма. Это неприятно, так как симметрия ведет к значительной экономии вычислений.

Если A_1 — положительно определенная, то мы можем модифицировать LR -алгоритм, используя симметричную факторизацию Холецкого (гл. 4, § 42). Обозначая так полученные матрицы через $\tilde{A}_s$, имеем

$$A_1 = L_1 L_1^T, \quad L_1^T L_1 = \tilde{A}_2 = L_1^{-1} A_1 L_1 = L_1^T A_1 (L_1^{-1})^T. \quad (50.2)$$

Очевидно, что $\tilde{A}_2$ симметрична и, так как она подобна A_1 , она также

положительно определенная. Следовательно, процесс можно продолжать, и можно доказать, что

$$\tilde{A}_s = L_{s-1}^{-1} \dots L_1^{-1} A_1 L_1 \dots L_{s-1} = L_{s-1}^T \dots L_1^T A_1 (L_1^{-1})^T \dots (L_{s-1}^{-1})^T \quad (50.3)$$

и

$$L_1 L_2 \dots L_s L_s^T \dots L_1^T = A_1^s \quad (50.4)$$

или

$$(L_1 L_2 \dots L_s) (L_1 L_2 \dots L_s)^T = A_1^s. \quad (50.5)$$

При этом не только вдвое уменьшается объем вычислений, но, как мы видели в гл. 4, §§ 43, 44, факторизация Холецкого обеспечивает высокую численную устойчивость и не требует перестановок. Если A_1 — неположительно определенная, то факторизация Холецкого приводит к комплексным числам, и численная устойчивость не обеспечивается. При этом A_2 , вообще говоря, даже неэрмитова (просто комплексная симметричная), но она, конечно, имеет вещественные собственные значения, так как подобна A_1 . Так, если

$$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{то} \quad \tilde{A}_2 = \begin{bmatrix} 2 & i \\ i & -1 \end{bmatrix}. \quad (50.6)$$

Связь между LR - и QR -алгоритмами

51. QR -алгоритм и LR -алгоритм Холецкого очень тесно связаны. На самом деле, имеем

$$(Q_1 Q_2 \dots Q_s) (R_s \dots R_2 R_1) = A_1^s, \quad (51.1)$$

и, следовательно,

$$(R_s \dots R_2 R_1)^T (Q_1 Q_2 \dots Q_s)^T (Q_1 Q_2 \dots Q_s) (R_s \dots R_2 R_1) = (A_1^s)^T A_1^s = A_1^{2s} \quad (51.2)$$

или

$$(R_s \dots R_2 R_1)^T (R_s \dots R_2 R_1) = A_1^{2s}.$$

Но из (50.5) следует, что

$$(L_1 L_2 \dots L_{2s}) (L_1 L_2 \dots L_{2s})^T = A_1^{2s}. \quad (51.3)$$

Следовательно, мы имеем две факторизации Холецкого матрицы A_1^{2s} , и так как они должны совпадать, находим

$$L_1 L_2 \dots L_{2s} = (R_s \dots R_2 R_1)^T. \quad (51.4)$$

Из (50.3) имеем

$$\tilde{A}_{2s+1} = (L_1 L_2 \dots L_{2s})^T A_1 [(L_1 L_2 \dots L_{2s})^T]^{-1}, \quad (51.5)$$

а из уравнений, определяющих QR -алгоритм, —

$$\begin{aligned} A_{s+1} &= R_s \dots R_1 A_1 (R_s \dots R_1)^{-1} = \\ &= (L_1 L_2 \dots L_{2s})^T A_1 [(L_1 L_2 \dots L_{2s})^T]^{-1} = \tilde{A}_{2s+1}. \end{aligned} \quad (51.6)$$

Следовательно, $(2s+1)$ -я матрица, полученная по LR -алгоритму Холецкого, равна $(s+1)$ -й матрице, полученной при использовании QR -алгоритма. (Этот результат несправедлив для несимметричных матриц, но он, возможно, демонстрирует большую силу QR .)

Наше доказательство справедливо как для положительно определенных матриц, так и для произвольных симметричных. (Конечно, мы требуем, чтобы они были неособенными, так как иначе треугольная факторизация A_1^s не единственна.)

QR-алгоритм всегда дает вещественные матрицы, если A_1 — вещественная, и, следовательно, LR-алгоритм Холецкого должен давать вещественные $\tilde{A}_{2s+1}$, даже если A_{2s} — комплексные. В случае матрицы (50.6) имеем

$$\tilde{A}_3 = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}. \quad (51.7)$$

Еще более важным является тот факт, что QR-алгоритм сохраняет 2-норму и евклидову норму, и ни один элемент A_s не может стать большим. Это неверно для LR-алгоритма Холецкого, если только A_1 не положительно определенная. Матрица $\tilde{A}_2$ может иметь сколь угодно большие элементы, но элементы $\tilde{A}_3$ обязательно должны вернуться к нормальным размерам. Пример неустойчивости приведен в табл. 4. Здесь $\|\tilde{A}_2\|_E$ значительно

Таблица 4

Вычисленные преобразования

$$\begin{aligned} A_1 &= \begin{bmatrix} 10^{-3} (0,4000) & 10^0 (0,9877) \\ 10^0 (0,9877) & 10^0 (0,1471) \end{bmatrix} & L_1 &= \begin{bmatrix} 10^{-1} (0,4000) & 0 \\ 10^2 (0,9877) & 10^2 (0,9877) i \end{bmatrix} \\ \tilde{A}_2 &= \begin{bmatrix} 10^4 (0,9756) & 10^4 (0,9756) i \\ 10^4 (0,9756) i & 10^4 (0,9756) \end{bmatrix} & L_2 &= \begin{bmatrix} 10^2 (0,9877) & 0 \\ 10^2 (0,9877) i & 10^0 (0,6979) i \end{bmatrix} \\ \tilde{A}_3 &= \begin{bmatrix} 0 & 10^2 (-0,6893) \\ 10^2 (-0,6893) & 10^0 (-0,4871) \end{bmatrix} \\ \text{Точная } \tilde{A}_3 &= \begin{bmatrix} 0,1473 & 0,9877 \\ 0,9877 & -0,0001 \end{bmatrix} \end{aligned}$$

больше $\|A_1\|_E$. Несмотря на то, что вычисленная $\tilde{A}_3$ — вещественная, произошла полная потеря точности собственных значений. Для сравнения дана точная $\tilde{A}_3$.

Сходимость LR-алгоритма Холецкого

52. Если A_1 — положительно определенная, то $\tilde{A}_s$ стремится к диагональной, независимо от природы собственных значений. Мы уже доказали, что QR-алгоритм для положительно определенных матриц всегда сходится, и связь LR-алгоритма Холецкого и QR-алгоритма обеспечивает сходимость первого. Следующее доказательство не зависит от этой связи.

Пусть A_{rs} и L_{rs} обозначают угловые главные миноры $\tilde{A}_s$ и L_s соответственно. Из соотношения $\tilde{A}_s = L_s L_s^T$ следует;

$$A_{rs} = (L_{rs})^2. \quad (52.1)$$

С другой стороны, из соотношения $\tilde{A}_{s+1} = L_s^T L_s$ и теоремы Бине — Коши $A_{r, s+1}$ равен сумме произведений соответствующих миноров порядка r , образованных из первых r строк L_s^T и первых r столбцов L_s . Но соответствующие миноры равны, так как эти матрицы транспонированы друг

другу, и, следовательно, $A_{r, s+1}$ представляется в виде суммы квадратов. Одно слагаемое равно L_{rs}^2 и, следовательно,

$$A_{r, s+1} \geq A_{rs}. \quad (52.2)$$

Поэтому каждый угловой минор $\tilde{A}_s$ увеличивается с ростом s , и в силу их очевидной ограниченности они стремятся к пределу. Из этого немедленно следует, что все недиагональные элементы стремятся к нулю. Методом, аналогичным §§ 33, 34, можно показать, что если собственные значения различны и предельная диагональная матрица имеет собственные значения, расположенные в убывающем порядке, то $\tilde{a}_{ij}^{(s)}$ стремятся к нулю как $(\lambda_j/\lambda_i)^{\frac{1}{2}s}$, когда $j > i$, и как $(\lambda_i/\lambda_j)^{\frac{1}{2}s}$ при $i > j$.

Интересно сравнить стандартный LR и LR Холецкого, когда они применяются к одной и той же симметричной матрице. Если $A_1 = \tilde{L}_1 \tilde{L}_1^T$ и $A_1 = L_1 R_1$, то, предполагая, что A_1 — неособенная, мы должны иметь

$$L_1 = \tilde{L}_1 D_1, \quad R_1 = D_1^{-1} \tilde{L}_1^T, \quad (52.3)$$

где D_1 — некоторая диагональная матрица. Следовательно,

$$\tilde{A}_2 = \tilde{L}_1^T \tilde{L}_1, \quad A_2 = R_1 L_1 = D_1^{-1} \tilde{L}_1^T \tilde{L}_1 D_1 = D_1^{-1} \tilde{A}_2 D_1. \quad (52.4)$$

Аналогично, если

$$\tilde{A}_2 = \tilde{L}_2 \tilde{L}_2^T, \quad A_2 = L_2 R_2 = D_1^{-1} \tilde{L}_2 \tilde{L}_2^T D_1, \quad (52.5)$$

то из единственности, с точностью до диагональной матрицы, разложения в произведение треугольных мы должны иметь

$$L_2 = D_1^{-1} \tilde{L}_2 D_2, \quad R_2 = D_2^{-1} \tilde{L}_2^T D_1 \quad (52.6)$$

для некоторой диагональной D_2 . Следовательно,

$$\tilde{A}_3 = \tilde{L}_2^T \tilde{L}_2, \quad A_3 = R_2 L_2 = D_2^{-1} \tilde{L}_2^T D_1 D_1^{-1} \tilde{L}_2 D_2 = D_2^{-1} \tilde{A}_3 D_2. \quad (52.7)$$

Мы видим, что на каждом этапе A_s получается из $\tilde{A}_s$ при помощи преобразования подобия с диагональной матрицей, так что

$$A_s = D_{s-1}^{-1} \tilde{A}_s D_{s-1}. \quad (52.8)$$

Заметим, что все это верно, даже если A_1 не положительно определенная, но в этом случае матрицы D_{s-1} будут комплексными при четных s , и может наступить стадия, на которой A_s не имеет разложения в произведение треугольных.

В положительно определенном случае мы знаем, что $\tilde{A}_s$ обязательно становятся диагональными, в то время как A_s становятся лишь верхними треугольными. Элементы D_s обязательно порядка $(\lambda_i)^{-\frac{1}{2}s}$.

53. Можно применить сдвиги в LR -алгоритме Холецкого, но если мы хотим, чтобы все элементы оставались вещественными и процесс был численно устойчивым, они должны быть выбраны так, чтобы $(A_s - k_s I)$ были положительно определенными на каждой стадии. С другой стороны, для быстрой сходимости желательно k_s выбирать как можно более близким к λ_n , наименьшему собственному значению. Таким образом, выбор сдвига является более деликатной проблемой, чем в случае матриц Хессенберга. Таких трудностей не возникает с QR -алгоритмом, где вообще не надо беспокоиться о положительной определенности.

Кубическая сходимость QR-алгоритма

54. Простейший выбор сдвига это $k_s = a_{nn}^{(s)}$. Мы сейчас покажем, что, вообще говоря, для QR-алгоритма это обязательно даст кубическую сходимость к наименьшему по модулю собственному значению, *независимо от его кратности*. Предположим, что

$$\left. \begin{aligned} \lambda_n &= \lambda_{n-1} = \dots = \lambda_{n-r+1}, \\ |\lambda_1| &\geq |\lambda_2| \geq \dots \geq |\lambda_{n-r}| > |\lambda_{n-r+1}|. \end{aligned} \right\} \quad (54.1)$$

Если мы положим

$$A_s = {}^{n-r} \left\{ \begin{array}{c|c} F_s & G_s \\ \hline G_s^T & H_s \end{array} \right\}, \quad (54.2)$$

то из общей теории § 32 мы знаем, что если не используются сдвиги, то G_s стремится к нулевой матрице, собственные значения λ'_i ($i = 1, \dots, n-r$) матрицы F_s стремятся к соответствующим λ_i , а все собственные значения λ''_i ($i = n-r+1, \dots, n$) матрицы H_s стремятся к λ_n . Предположим, что мы достигли стадии, когда

$$A_s = \left[\begin{array}{c|c} F_s & \varepsilon K_s \\ \hline \varepsilon K_s^T & H_s \end{array} \right], \quad \|K_s\|_E = 1, \quad (54.3)$$

$$|\lambda'_i - \lambda_n| > \frac{2}{3} |\lambda_{n-r} - \lambda_n| \quad (i = 1, \dots, n-r), \quad \varepsilon < \frac{1}{3} |\lambda_{n-r} - \lambda_n|. \quad (54.4)$$

Мы знаем, что ранг $(A_s - \lambda_n I)$ равен $n-r$ на всех шагах. Из (54.4) следует, что ранг $(F_s - \lambda_n I)$ также равен $n-r$. Следовательно, мы должны иметь

$$H_s - \lambda_n I - \varepsilon^2 K_s^T (F_s - \lambda_n I)^{-1} K_s = 0, \quad (54.5)$$

что дает

$$H_s = \lambda_n I + \varepsilon^2 K_s^T (F_s - \lambda_n I)^{-1} K_s = \lambda_n I + M_s. \quad (54.6)$$

Имеем

$$\begin{aligned} \|M_s\|_E &\leq \varepsilon^2 \|K_s\|_E^2 \|(F_s - \lambda_n I)^{-1}\|_2 \leq \\ &\leq \varepsilon^2 / \frac{2}{3} |\lambda_{n-r} - \lambda_n| = \frac{3}{2} \varepsilon^2 / |\lambda_{n-r} - \lambda_n| < \frac{1}{6} |\lambda_{n-r} - \lambda_n|. \end{aligned} \quad (54.7)$$

Эти результаты показывают, что норма недиагональных элементов H_s порядка ε^2 , и что все диагональные элементы H_s отличаются от λ_n на величины порядка ε^2 . В частности,

$$|a_{nn}^{(s)} - \lambda_n| < \frac{3}{2} \varepsilon^2 / |\lambda_{n-r} - \lambda_n| < \frac{1}{6} |\lambda_{n-r} - \lambda_n|, \quad (54.8)$$

и мы выводим, что

$$|\lambda'_i - a_{nn}^{(s)}| > |\lambda'_i - \lambda_n| - |\lambda_n - a_{nn}^{(s)}| > \frac{1}{2} |\lambda_{n-r} - \lambda_n| \quad (54.9)$$

$$(i = 1, \dots, n-r),$$

$$|\lambda''_i - a_{nn}^{(s)}| < |\lambda''_i - \lambda_n| + |\lambda_n - a_{nn}^{(s)}| < 3\varepsilon^2 / |\lambda_{n-r} - \lambda_n| \quad (54.10)$$

$$(i = n-r+1, \dots, n).$$

Рассмотрим теперь эффект выполнения одного шага QR со сдвигом $a_{nn}^{(s)}$. Так как $\lambda_n - a_{nn}^{(s)}$ порядка ε^2 , а $\lambda_i - a_{nn}^{(s)}$ ($i = 1, \dots, n - r$) не зависят от ε , то из общей теории сходимости (§ 19) можно ожидать, что элементы εK_s^T будут умножаться на множитель порядка ε^2 , но полезно дать точное доказательство. Пусть P_s это такая ортогональная матрица, что $P_s (A_s - a_{nn}^{(s)} I)$ — верхняя треугольная. Тогда, написав

$$P_s = \left[\begin{array}{c|c} Q_s & R_s \\ \hline S_s & T_s \end{array} \right], \quad A_{s+1} = P_s (A_s - a_{nn}^{(s)} I) P_s^T + a_{nn}^{(s)} I =$$

$$= \left[\begin{array}{c|c} F_{s+1} & G_{s+1} \\ \hline G_{s+1}^T & H_{s+1} \end{array} \right], \quad (54.11)$$

имеем

$$0 = S_s (F_s - a_{nn}^{(s)} I) + \varepsilon T_s K_s^T \quad (54.12)$$

и

$$G_{s+1}^T = [\varepsilon S_s K_s + T_s (H_s - a_{nn}^{(s)} I)] R_s^T. \quad (54.13)$$

Следовательно,

$$\|G_{s+1}\|_E \leq \varepsilon \|S_s\|_E^2 + \|T_s\|_E \|S_s\|_E \|H_s - a_{nn}^{(s)} I\|_2 \quad (\|R_s\|_E = \|S_s\|_E), \quad (54.14)$$

и из (54.12) и (54.10)

$$\|S_s\|_E = \varepsilon \|T_s K_s^T (F_s - a_{nn}^{(s)} I)^{-1}\|_E \leq$$

$$\leq \varepsilon \|T_s\|_E \max |\lambda_i' - a_{nn}^{(s)}|^{-1} < 2\varepsilon \|T_s\|_E / |\lambda_{n-r} - \lambda_n|. \quad (54.15)$$

Следовательно, окончательно имеем

$$\|G_{s+1}\|_E \leq 4\varepsilon^3 \|T_s\|_E^2 / |\lambda_{n-r} - \lambda_n|^2 + 2\varepsilon \|T_s\|_E^2 \max |\lambda_i' - a_{nn}^{(s)}| / |\lambda_{n-r} - \lambda_n| <$$

$$< 10\varepsilon^3 \|T_s\|_E^2 / |\lambda_{n-r} - \lambda_n|^2 < 10r\varepsilon^3 / |\lambda_{n-r} - \lambda_n|^2. \quad (54.16)$$

Таким образом, кубическая сходимость доказана. Заметим, что это означает, что все r совпадающих собственных значений будут найдены одновременно, так что совпадение собственных значений является скорее преимуществом, чем недостатком.

На практике мы начинаем применять сдвиг значительно раньше, чем в нашем доказательстве, обычно тогда, когда установился один двоичный знак $a_{nn}^{(s)}$.

Сдвиг в LR -алгоритме Холецкого

55. Возвращаясь к LR -алгоритму Холецкого, замечаем, что если

$$(A_s - k_s I) = L_s L_s^T, \quad (55.1)$$

то по (50.3)

$$\det(A_1 - k_s I) = \det(A_s - k_s I) = \left(\prod l_{ii}^{(s)} \right)^2. \quad (55.2)$$

Следовательно, каждый раз, когда выполняется факторизация, мы можем вычислить $\det(A_1 - \lambda I)$ при $\lambda = k_s$. Если k_s и k_{s-1} меньше λ_n , то в силу того, что

$$\det(A_1 - \lambda I) = f(\lambda) = \prod (\lambda_i - \lambda), \quad (55.3)$$

при

$$k_{s+1} = [f(k_s)k_{s-1} - f(k_{s-1})k_s] / [f(k_s) - f(k_{s-1})] \quad (55.4)$$

получим, что

$$k_{s+1} < \lambda_n, \quad \lambda_n - k_{s+1} < (\lambda_n - k_s) \quad \text{и} \quad (\lambda_n - k_{s-1}). \quad (55.5)$$

Как только мы смогли выбрать k_1 и k_2 , меньшие λ_n , мы сможем построить и последовательность таких значений, каждое из которых меньше λ_n и которые сходятся к нему. Если A_1 — положительно определенная, то можем взять $k_1 = -2^{-\frac{1}{2}t} \|A_1\|$, $k_2 = 0$. Все следующие k_i будут положительными. Заметим, что мы не должны продолжать итерации до тех пор, пока k_s станет с рабочей точностью равным λ_n . Мы можем остановиться, когда недиагональные элементы n -й строки и n -го столбца станут пренебрежимо малыми, и провести исчерпывание. Если мы храним два последних значения $\prod_{i=1}^{n-1} l_{ii}^{(s)}$ на всех стадиях, то можем вычислить хорошие начальные значения для сдвига сразу после исчерпывания.

Эта техника довольно проста и очень эффективна, если нет кратных или патологически близких собственных значений.

Срыв факторизации Холецкого

56. Рутисхаузер (1960) предложил способ выбора сдвигов, который обязательно дает кубическую сходимость. Он основан на анализе срыва, который случается, если берется k_s большее, чем λ_n . Если мы напишем, опуская для простоты индекс s ,

$$A - kI = LL^T, \quad A = {}^P \left[\begin{array}{c|c} F & G \\ \hline G^T & H \end{array} \right], \quad L = \left[\begin{array}{c|c} M & 0 \\ \hline N & P \end{array} \right], \quad (56.1)$$

то

$$MM^T = F - kI, \quad MN^T = G, \quad NN^T + PP^T = H - kI. \quad (56.2)$$

Предположим, что мы достигли такой стадии, когда M и N уже определены, но при попытке определить первый элемент P нужно извлекать квадратный корень из отрицательной величины. Из (56.2) имеем

$$PP^T = H - kI - G^T(F - kI)^{-1}G = X, \quad (56.3)$$

и, по нашему предположению, x_{11} отрицателен. Предположим, что τ — алгебраически наименьшее собственное значение X . Из наших предположений о X оно, очевидно, отрицательно. Утверждаем, что $(A - kI - \tau I)$ — положительно определенная. Другими словами, если мы используем сдвиг $(k + \tau)$ вместо k , то сможем продолжать наш процесс.

Очевидно, что так как τ отрицательно, мы сможем достичь по крайней мере той же стадии, что раньше, и соответствующая P теперь определяется соотношением

$$\begin{aligned} PP^T &= H - kI - \tau I - G^T(F - kI - \tau I)^{-1}G = \\ &= (X - \tau I) + G^T(F - kI)^{-1}G - G^T(F - kI - \tau I)^{-1}G = Y. \end{aligned} \quad (56.4)$$

Мы должны показать, что Y — положительно определенная. Так как τ — наименьшее собственное значение X , $(X - \tau I)$ — неотрицательно

определенная. Пусть Q — такая ортогональная матрица, что

$$F = Q^T \text{diag}(\lambda_i) Q. \quad (56.5)$$

Так как мы считаем, что $(F - kI)$ имеет факторизацию, то $\lambda'_i - k > 0$ ($i = 1, \dots, p$). Следовательно, имеем

$$\begin{aligned} G^T (F - kI)^{-1} G - G^T (F - kI - \tau I)^{-1} G &= \\ = G^T Q^T [\text{diag}(\lambda'_i - k)^{-1} - \text{diag}(\lambda'_i - k - \tau)^{-1}] Q G &= \\ = S^T \text{diag}[-\tau/(\lambda'_i - k)(\lambda'_i - k - \tau)] S. \end{aligned} \quad (56.6)$$

Все элементы диагональной матрицы положительны, и, следовательно, матрица в правой части (56.6), конечно, неотрицательно определенная, и она положительно определенная, если только S и, следовательно, G не имеют линейно зависимых строк. Поэтому Y всегда неотрицательно определенная и, вообще говоря, положительно определенная.

Для получения пользы из этого результата существенно, во-первых, чтобы срыв произошел на очень поздней стадии, так чтобы X была матрицей низкого порядка (предпочтительно 1 или 2) и ее собственные значения можно было легко найти, и, во-вторых, чтобы $k + \tau$ было очень близко к λ_n . Мы рассмотрим эту проблему в следующем параграфе.

Кубически сходящийся LR-процесс

57. Пусть A_1 — матрица типа, который мы обсуждали в § 54. Кроме того, предположим, что она положительно определенная и что мы получили A_s , удовлетворяющую условиям (54.3), (54.4), используя LR-алгоритм Холецкого. Из (54.6) знаем, что на этой стадии все диагональные элементы H_s отличаются от λ_n на величины порядка ε^2 . Предположим, что мы используем в качестве сдвига $a_{nn}^{(s)} = \lambda_n + m_{nn}^{(s)}$. При этом факторизация должна сорваться, так как все диагональные элементы симметричной матрицы не меньше ее наименьшего собственного значения. Однако из (54.9) следует, что срыв не произойдет ранее стадии $(n - r + 1)$. Матрица X на этой стадии имеет вид

$$\begin{aligned} X &= H_s - (\lambda_n + m_{nn}^{(s)}) I - \varepsilon^2 K_s^T (F_s - a_{nn}^{(s)} I)^{-1} K_s = \\ &= \lambda_n I + \varepsilon^2 K_s^T (F_s - \lambda_n I)^{-1} K_s - (\lambda_n + m_{nn}^{(s)}) I - \varepsilon^2 K_s^T (F_s - a_{nn}^{(s)} I)^{-1} K_s = \\ &= -m_{nn}^{(s)} I + \varepsilon^2 K_s^T [(F_s - \lambda_n I)^{-1} - (F_s - a_{nn}^{(s)} I)^{-1}] K_s, \end{aligned} \quad (57.1)$$

и эта матрица порядка r . Если τ — наименьшее собственное значение X , то (см. § 56) $a_{nn}^{(s)} + \tau$ является подходящим сдвигом. Мы покажем, что это очень близко к λ_n . Положив $X = -m_{nn}^{(s)} I + Y$, получим

$$\|Y\|_E \leq \varepsilon^2 \|(F_s - \lambda_n I)^{-1} - (F_s - a_{nn}^{(s)} I)^{-1}\|_2. \quad (57.2)$$

Так как обе матрицы в правой части имеют одну систему собственных векторов, то, используя (54.8), (54.9) и (54.4), имеем

$$\begin{aligned} \|Y\|_E &\leq \varepsilon^2 \max |1/(\lambda'_i - \lambda_n) - 1/(\lambda'_i - a_{nn}^{(s)})| = \\ &= \varepsilon^2 \max |(a_{nn}^{(s)} - \lambda_n)/(\lambda'_i - \lambda_n)(\lambda'_i - a_{nn}^{(s)})| \leq \\ &\leq \varepsilon^2 \left[\frac{3}{2} \varepsilon^2 / (\lambda_{n-r} - \lambda_n) \right] [3/(\lambda_{n-r} - \lambda_n)^2] = \frac{9}{2} \varepsilon^4 / (\lambda_{n-r} - \lambda_n)^3. \end{aligned} \quad (57.3)$$

Следовательно, так как $\tau = -m_{nn}^{(s)} +$ (наименьшее собственное значение Y), имеем

$$a_{nn}^{(s)} + \tau = \lambda_n + m_{nn}^{(s)} - m_{nn}^{(s)} + (\text{наименьшее собственное значение } Y), \quad (57.4)$$

что дает

$$|a_{nn}^{(s)} + \tau - \lambda_n| \leq \|Y\|_E \leq \frac{9}{2} \varepsilon^4 / (\lambda_{n-r} - \lambda_n)^3. \quad (57.5)$$

Для того чтобы показать, что после итерации с матрицей $A_s - (a_{nn}^{(s)} + \tau) I$ евклидова норма матрицы G_{s+1} имеет порядок ε^3 , мы могли бы провести прямой анализ, аналогичный анализу § 54. Оставим это для упражнения. Общая теория показывает, что G_s умножится на множитель порядка

$$|[\lambda_n - (a_{nn}^{(s)} + \tau)][\lambda_{n-r} - (a_{nn}^{(s)} + \tau)]^{1/2}|,$$

т. е. порядка ε^2 . Так как G_s порядка ε , результат установлен.

58. Мы рассмотрели случай, когда кратность наименьшего собственного значения произвольна, но на практике метод наиболее удовлетворителен, если r равно 1 или 2, так как только тогда легко определяется наименьшее собственное значение τ .

Довольно трудно определить, достигли ли мы той стадии, когда полезно применять вышеописанную процедуру, и на практике решение должно быть принято из эмпирических соображений. Общая стратегия для простого собственного значения заключается в следующем.

Определяем положительные сдвиги любым подходящим способом, например, как в § 55, до тех пор, пока $(a_{nn}^{(s)} - k_{s-1})$ не станет значительно меньше всех остальных диагональных элементов A_s и внедиагональные элементы в последних столбце и строке не станут малыми. Затем в качестве следующего сдвига берем $a_{nn}^{(s)}$ и разложение не проходит. Если при этом порядок матрицы, остающейся ниже точки срыва, больше двух, то мы сменили методы слишком рано. (Заметим, что если кратность собственного значения действительно больше двух, то срыв обязательно произойдет слишком рано.) Рутисхаузер и Шварц (1963) развили очень изобретательные стратегии для приложения этого процесса, и мы советуем читателю обратиться к их работе.

К сожалению, надо заметить, что теоретически проще использовать «невосстанавливающий» метод LR Холецкого. Тогда элемент $a_{nn}^{(s)}$ дает важную информацию о сходимости. Так как A_s образуется вычислением $L_{s-1}^T L_{s-1}$, она, по крайней мере, неотрицательно определенная. Если $a_{nn}^{(s)} < \varepsilon$, то исчерпывание на этой стадии даст ошибку в каждом собственном значении, меньшую ε . Действительно, если собственные значения ведущей главной подматрицы порядка $(n-1)$ равны λ'_i , то

$$\left. \begin{aligned} \lambda_1 &> \lambda'_1 > \dots > \lambda_{n-1} > \lambda'_{n-1}, \\ \lambda'_1 + \dots + \lambda'_{n-1} + \varepsilon &= \lambda_1 + \dots + \lambda_n, \\ (\lambda_1 - \lambda) + \dots + (\lambda_{n-1} - \lambda'_{n-1}) &= \varepsilon - \lambda_n > 0. \end{aligned} \right\} \quad (58.1)$$

Отсюда следует результат. Заметим, что если недиагональные элементы последних строки и столбца достаточно малы, то мы можем произвести исчерпывание независимо от того, мал или нет $a_{nn}^{(s)}$. Этот элемент будет на самом деле мал, если только сдвиг был выбран хорошо.

Ленточные матрицы

59. Объем работы для произвольных симметричных матриц очень велик, но это не так для ленточных матриц, т. е. матриц таких, что

$$a_{ij} = 0 \quad (|i - j| > m), \quad 2m + 1 \ll n. \quad (59.1)$$

Рассмотрим сначала LR -алгоритм Холецкого. Очевидно, что он сохраняет ленточную форму (для несимметричных матриц ср. § 66), и на каждом шаге матрица L_s такая, что

$$l_{ij}^{(s)} = 0 \quad (i - j > m). \quad (59.2)$$

Следовательно, хотя A_s имеет $2m + 1$ отличных от нуля элементов в типичной строке, L_s имеет только $m + 1$ элементов.

Существует много путей выполнения разложения и организации запоминания. Следующий способ при условии, что желательно получить все преимущества от накопления скалярных произведений, наверное, наиболее удобный.

Для того чтобы получить все преимущества от ленточной формы и симметрии, удобно обозначить элементы A следующим образом, проиллюстрированным для случая $n = 8$, $m = 3$.

Верхний треугольник A

a_{10}	a_{11}	a_{12}	a_{13}
	a_{20}	a_{21}	a_{22}
		a_{30}	a_{31}
			a_{40}

Заметим, что A записывается как прямоугольник $n \times (m + 1)$. Нижний правый угол A заполнен нулями, но эти элементы могут быть произвольными, если разложение будет выполняться так, как описано ниже. Пунктирные линии столбцы A . Верхняя треугольная $U = L^T$ получается в той же форме, что и верхняя половина A . Она может быть записана после вычисления на месте A , но если использованный сдвиг был больше λ_n , этого делать нельзя, так как в этом случае разложение сорвется, и нужно будет начать процесс снова.

60. Элементы U получаются строка за строкой, причем элементы i -й строки вычисляются следующим образом:

(i) Определяем $p = \min(n - i, m)$.

Затем для всех j от 0 до p выполняем шаги (ii) и (iii).

(ii) Определяем $q = \min(m - j, i - 1)$.

(iii) Вычисляем $x = a_{ij} - \sum_{k=1}^q (u_{i-k, k})(u_{i-k, j+k})$, где сумма равна нулю, если $q = 0$.

Если $j = 0$, то $u_{i0} = (x - k_s)^{1/2}$. (Если $x - k_s$ отрицательно, то k_s слишком велик.)

Если $j \neq 0$, то $u_{ij} = x/u_{i0}$.

В качестве примера при $i = 5$, $m = 3$, $n = 8$ имеем

$$\left. \begin{aligned} u_{50}^2 &= a_{50} - u_{41}u_{41} - u_{32}u_{32} - u_{23}u_{23} - k_s, \\ u_{50}u_{51} &= a_{51} - u_{41}u_{42} - u_{32}u_{33}, \\ u_{50}u_{52} &= a_{52} - u_{41}u_{43}, \\ u_{50}u_{53} &= a_{53}. \end{aligned} \right\} \quad (60.1)$$

Величины p и q введены для того, чтобы учесть эффекты окончания.

Затем нужно вычислить $(UU^T + K_s I)$. Вычисленная матрица записывается на месте исходной A . Если U уже была записана на месте исходной A , то это не вызывает никаких трудностей, так как к моменту, когда a_{ij} уже вычислены, u_{ij} больше не нужны. Элементы i -й строки вычисляются следующим образом:

(i) Определяем $p = \min(n - i, m)$.

Затем для всех j от 0 до p выполняем шаги (ii) и (iii).

(ii) Определяем $q = \min(m - j, n - i - j)$.

(iii) Вычисляем $x = \sum_{k=0}^q u_{i, j+k} u_{i+j, k}$.

Если $j = 0$, то $a_{i0} = x + k_s$.

Если $j \neq 0$, то $a_{ij} = x$. Например, если $i = 5$, $m = 3$, то

$$\left. \begin{aligned} a_{50} &= u_{50}u_{50} + u_{51}u_{51} + u_{52}u_{52} + u_{53}u_{53} + k_s, \\ a_{51} &= u_{51}u_{60} + u_{52}u_{61} + u_{53}u_{62}, \\ a_{52} &= u_{52}u_{70} + u_{53}u_{71}, \\ a_{53} &= u_{53}u_{80}. \end{aligned} \right\} \quad (60.2)$$

На вычислительных машинах с двумя типами памяти разложение и перемножение *) могут быть скомбинированы. При вычислении строки i матрицы U_s строки $(i - 1)$, $(i - 2)$, ..., $(i - m)$ матрицы U_s и строка i матрицы A_s должны быть в быстродействующей памяти. После того как вычислена i -я строка U_s , можно вычислить $(i - m)$ -ю строку A_{s+1} . Снова мы не можем сделать это, если используется такое значение k_s , которое вызовет срыв.

61. Если используется техника Рутисхаузера § 56, то мы должны исследовать каждый срыв. Если срыв произошел, когда $i = n$, то текущее значение x является следующим сдвигом. Если срыв произошел, когда $i = n - 1$, мы должны вычислить матрицу X второго порядка из уравнения (56.3). Имеем

$$\left. \begin{aligned} x_{11} &= \text{текущее значение } x, \\ x_{12} &= a_{n-1, 1} - u_{n-2, 1}u_{n-2, 2} - u_{n-3, 2}u_{n-3, 3} - \dots - u_{n-m, m-1}u_{n-m, m}, \\ x_{22} &= a_{n, 0} - u_{n-2, 2}^2 - u_{n-3, 3}^2 - \dots - u_{n-m, m}^2, \end{aligned} \right\} \quad (61.1)$$

*) Так мы переводим термин автора recombination. (Прим. перев.)

и нам нужно наименьшее собственное значение этой матрицы второго порядка. Если срыв произошел на более ранней стадии, то не принято использовать оставшуюся матрицу.

Требуется примерно $\frac{1}{2}nm^2$ умножений для выполнения разложения и столько же для переумножения. Если $m \ll n$, это значительно меньше, чем для полной матрицы. Мы видели в гл. 4, § 40, что ошибки при разложении очень малы. Например, при использовании арифметики с фиксированной запятой с накоплением скалярных произведений имеем

$$L_s L_s^T = A_s + E_s, \quad |e_{ij}^{(s)}| \leq \frac{1}{2} 2^{-t}, \quad (61.2)$$

где E_s имеет ту же ленточную форму.

Аналогично

$$A_{s+1} = L_s^T L_s + F_s, \quad |f_{ij}^{(s)}| \leq \frac{1}{2} 2^{-t}, \quad (61.3)$$

где F_s снова имеет ленточную форму. Из этих результатов легко видеть, что максимальная ошибка, введенная в любое собственное значение полным преобразованием, равна $(2m + 1) 2^{-t}$. Если A_1 нормирована так, что $\|A_1\|_\infty < 1$, то все элементы всех A_s ограничены единицей. Заметим, что если мы сравним вычисленную A_s с точной A_s , нам может неправильно показаться, что собственные значения имеют значительно меньшую точность (см., например, Рутисхаузер (1958, стр. 80)).

QR-разложение ленточной матрицы

62. Значительно менее очевидно, что ленточная форма сохраняется в QR-алгоритме. Матрица R , получающаяся при QR-разложении, имеет ширину $2m + 1$, а не $m + 1$, как было в случае LR-алгоритма Холецкого. Если мы используем для приведения к треугольному виду метод Хаусхолдера (гл. 4, § 46), то необходимы $n - 1$ матрицы отражения $P_1, P_2, \dots, P_{n-1}$, причем элементы вектора w_i , соответствующего P_i , отличны от нуля только в позициях $i, i + 1, \dots, i + m$ (исключая, конечно, последние w_i). Матрицы $P_2 P_1 A_s$ и R_s в случае $n = 8, m = 2$ имеют вид

$$\begin{array}{c} P_2 P_1 A_s \\ \left[\begin{array}{cccccc} \times & \times & \times & \times & \times & \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ & 0 & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times & \times \\ & & & & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times & \times \end{array} \right], \end{array} \quad \begin{array}{c} P_{n-1} \dots P_1 A_s = R_s \\ \left[\begin{array}{cccccc} \times & \times & \times & \times & \times & \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times & \times \\ & & & & & & & \times & \times \\ & & & & & & & & \times \end{array} \right]. \end{array} \quad (62.1)$$

В первой показаны нули, полученные при помощи P_1 и P_2 . Рассмотрим теперь умножение справа на P_1 . Оно заменит столбцы от 1 до $m + 1$ на их

линейные комбинации. Следовательно, $R_s P_1$ имеет вид

$$\begin{array}{c} R_s P_1 \qquad \qquad \qquad R_s P_1 P_2 P_3 \\ \left[\begin{array}{cccccc} \times & \times & \times & \times & \times & \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times \\ & & & & \times & \times & \times \\ & & & & & \times & \times \\ & & & & & & \times \end{array} \right], \quad \left[\begin{array}{cccccc} \times & \times & \times & \times & \times & \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times & \times & \times \\ & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \times & \times \\ & & & & & & & \times \end{array} \right]. \quad (62.2)$$

Дальнейшее умножение справа на $P_2, \dots, P_{n-1}$ не меняет первого столбца, и, следовательно, A_{s+1} имеет в этом столбце $m+1$ элемент. Продолжая таким же образом далее, видим, что A_{s+1} имеет только m элементов ниже диагонали в каждом столбце (исключая, конечно, последние m столбцов). Однако мы знаем, что окончательная матрица симметрична, и следовательно, элементы в позициях i, j , где $j > i + m$ должны аннулироваться.

Очевидно, что для вычисления *нижнего треугольника* матрицы A_{s+1} мы используем только элементы R_s в позициях (i, j) , при $j = i, i+1, \dots, i+m$. Следовательно, при вычислении R_s нужно получать только первые $m+1$ ненулевых элементов в каждой строке.

63. Для использования этого факта и симметрии потребуется дополнительный объем памяти. Для простоты предположим сначала, что вычисление R_s завершено прежде, чем мы начинаем вычислять A_{s+1} . Приведение к треугольному виду наиболее просто описывается, если рассмотреть один шаг. Поэтому рассмотрим четвертый основной шаг в приведении к треугольному виду матрицы A_s в случае $n=8, m=2$. Состояние в начале этого шага будет таким:

$$\begin{array}{c} A_s \text{ и } R_s \qquad \qquad \qquad W_s \\ \left[\begin{array}{ccc} r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \\ r_{30} & r_{31} & r_{32} \\ a_{40} & a_{41} & a_{42} \\ a_{50} & a_{51} & a_{52} \\ a_{60} & a_{61} & a_{62} \\ a_{70} & a_{71} & \times \\ a_{80} & \times & \times \end{array} \right], \quad \left[\begin{array}{ccc} w_{10} & w_{11} & w_{12} \\ w_{20} & w_{21} & w_{22} \\ w_{30} & w_{31} & w_{32} \end{array} \right], \quad (63.1) \\ \text{Дополнительный объем памяти} \\ \left[\begin{array}{ccccc} y_{40} & y_{41} & y_{42} & y_{43} & 0 \\ z_{50} & z_{51} & z_{52} & z_{53} & 0 \\ a_{42} & a_{51} & (a_{60} - k_s) & a_{61} & a_{62} \end{array} \right]. \end{array}$$

Матрица W_s состоит из векторов w_i матриц P_i . В дополнительной памяти хранятся в общем случае $m+1$ векторов с $2m+1$ элементами в каждом. В нашем случае перед началом четвертого основного шага в ней уже содержатся векторы-строки, обозначенные $y_{40}, \dots, y_{43}$ и $z_{50}, \dots, z_{53}$, которые представляют собой частично обработанные строки 4 и 5. Расположение (63.1) такое, какое получилось бы при неприведенном к треугольному виду правом нижнем углу полной матрицы. Индексы векторов y и z обозначают различные расположения относительно диагонали. Четвертый шаг состоит в следующем:

(i) Переносим *полную* строку 6 матрицы A_s в дополнительную память и применяем сдвиг, как показано. (Заметим, что все необходимые строки A_s имеются в распоряжении.)

(ii) По элементам y_{40} , z_{50} , a_{42} вычисляем вектор w_4 . (В общем случае w_4 имеет $(m+1)$ ненулевых элементов.) Записываем w_4 как четвертую строку w_s .

(iii) Умножаем слева на P_4 . При этом меняются только элементы, записанные в дополнительной памяти. Эффект этого умножения (но не распределения памяти) показан в (63.2). Нужно вычислять только первые три элемента четвертой строки R_s

$$P_4 \times \begin{bmatrix} y_{40} & y_{41} & y_{42} & y_{43} & 0 \\ z_{50} & z_{51} & z_{52} & z_{53} & 0 \\ a_{42} & a_{51} & (a_{60} - k_s) & a_{61} & a_{62} \end{bmatrix} = \begin{bmatrix} r_{40} & r_{41} & r_{42} & - & - \\ 0 & y_{50} & y_{51} & y_{52} & y_{53} \\ 0 & z_{60} & z_{61} & z_{62} & z_{63} \end{bmatrix}. \quad (63.2)$$

По мере вычисления четвертой строки R_s , она записывается на месте четвертой строки A_s (заметим, что вычислены три первые компоненты), и по мере того, как вычисляются элементы y_{5i} и z_{6i} , они записываются в позициях для y и z в дополнительной памяти. Следовательно, после завершения (iii) все готово для проведения пятого основного шага. Конечно, здесь есть обычные эффекты окончания. Так как использование неудачных разложений при QR-алгоритме исключено, мы всегда можем записать R_s на место A_s .

64. Рассматривая переумножение, видим, что A_{s+1} вычисляется строка за строкой и записывается на место R_s . На самом деле мы вычисляем поддиагональные элементы каждого столбца A_{s+1} и затем представляем их как наддиагональные элементы соответствующей строки A_{s+1} , используя тем самым известную симметрию A_{s+1} и сильно экономя в вычислениях.

Опишем основной шаг, в котором будет вычислена четвертая строка A_{s+1} . Дополнительная память в общем случае содержит массив $(m+1) \times (m+1)$ ячеек. В нашем случае перед началом четвертого основного шага векторы-столбцы $\tilde{y}_{40}$, $\tilde{y}_{41}$ и $\tilde{z}_{50}$, $\tilde{z}_{51}$ (волна означает, что эти величины не связаны с величинами в (63.1)) уже находятся в дополнительной памяти, как показано в (64.1):

$$\begin{bmatrix} \tilde{y}_{40} & \tilde{z}_{50} & r_{42} \\ \tilde{y}_{41} & \tilde{z}_{51} & r_{51} \\ 0 & 0 & r_{60} \end{bmatrix} \times 'P_4' = \begin{bmatrix} a_{40} & - & - \\ a_{41} & \tilde{y}_{50} & \tilde{z}_{60} \\ a_{42} & \tilde{y}_{51} & \tilde{z}_{61} \end{bmatrix}. \quad (64.1)$$

Четвертый основной шаг состоит в следующем:

(i) Переносим элементы r_{42} , r_{51} , r_{60} шестого столбца R_s в дополнительную память, как показано.

(ii) Умножаем справа матрицу третьего порядка в дополнительной памяти на $'P_4'$, подматрицу P_4 , которая является функцией w_4 . Первый столбец полученной матрицы дает соответствующие элементы четвертой строки A_{s+1} . Два остальных столбца дают векторы $\tilde{y}$ и $\tilde{z}$ для следующего основного шага. Заметим, что на самом деле выполняется только часть вычислений, связанных с умножением справа на P_4 .

Теперь очевидно, что разложение и переумножение могут быть скомбинированы. В нашем примере после вычисления первых трех строк R_s имеется достаточно информации для определения первой строки A_{s+1} . В общем случае разложение должно на m шагов опережать переумножение, и если будем делать так, надо будет запоминать не всю W_s , а только

последние m векторов w_i . Так как мы рекомендуем использовать QR- (и LR-) алгоритм только для узких ленточных матриц, дополнительная память мала по сравнению с памятью для A_s . Интересно, что если комбинируем разложение и перемножение, то объем памяти в QR меньше, чем в LR Холецкого, в котором используются неудачные разложения!

Анализ ошибок

65. Используя симметрию, мы предполагаем, что элементы a_{ij} ($m < j - i \leq 2m$), которые были бы равны нулю при точных вычислениях, на самом деле будут пренебрежимо малыми. Важно оправдать это предположение, и, к счастью, это почти следует из нашего общего анализа § 45 гл. 3. Если $\bar{R}_s$ — вычисленная матрица (с точными нулями под диагональю), а $P_1, P_2, \dots, P_{n-1}$ — точные матрицы отражения, соответствующие последовательным вычисленным приведенным матрицам, то имеем

$$\bar{R}_s = (P_{n-1}P_{n-2} \dots P_1) A_s + E_s, \quad (65.1)$$

где E_s мала. На самом деле, вычисляя только часть $\bar{R}_s$, мы используем в точности те же арифметические операции, какие использовались бы при вычислении всей этой матрицы.

Умножим теперь вычисленную $\bar{R}_s$ последовательно на вычисленные матрицы отражения $\bar{P}_r$. Несмотря на все возможные ошибки округления, результирующая матрица $\bar{A}_{s+1}$ наверняка имеет только m поддиагональных элементов в каждом столбце. При условии, что были проведены полные вычисления, общий анализ показывает, что $\bar{A}_{s+1}$ удовлетворяет соотношению

$$\begin{aligned} \bar{A}_{s+1} &= \bar{R}_s P_1 \dots P_{n-1} + F_s = \\ &= P_{n-1} \dots P_1 A_s P_1 \dots P_{n-1} + E_s P_1 \dots P_{n-1} + F_s, \end{aligned} \quad (65.2)$$

где F_s мала. На самом же деле мы таким образом вычисляем только нижнюю половину $\bar{A}_{s+1}$, и по симметрии получаем матрицу $\bar{\bar{A}}_{s+1}$.

Очевидно, имеем

$$\begin{aligned} \|G_{s+1}\|_E &= \|\bar{\bar{A}}_{s+1} - P_{n-1} \dots P_1 A_s P_1 \dots P_{n-1}\|_E \leq \\ &\leq 2^{1/2} [\|E_s P_1 \dots P_{n-1}\|_E + \|F_s\|_E] = 2^{1/2} [\|E_s\|_E + \|F_s\|_E]. \end{aligned} \quad (65.3)$$

Для вычислений с плавающей запятой и накоплением скалярных произведений можно показать, что

$$\|G_{s+1}\|_E \leq Km 2^{-t} \|A_1\|_E \quad (65.4)$$

с некоторой постоянной K , так что можно ожидать очень высокую точность.

На одном шаге QR приблизительно в три раза больше работы, чем на одном шаге LR Холецкого, хотя мы видели (§ 51), что если не используется сдвиг, один шаг QR эквивалентен двум шагам LR Холецкого. Выбор сдвига значительно проще для QR. Если брать $k_s = a_{nn}^{(s)}$, то сходимость обязательно будет кубическая. Если брать k_s равным наименьшему собственному значению матрицы второго порядка из нижнего правого угла, то получим даже лучшую сходимость. Обычно на практике требуется очень мало итераций. Нет гарантии, что собственные значения будут

получены в возрастающем порядке, хотя если исходная матрица положительно определенная и выполнены один или два предварительных шага с нулевым сдвигом, то это обычно верно.

Выбор сдвига в LR Холецкого значительно сложнее, если используется техника Рунтисхаузера с неудачными разложениями, но тем не менее эта схема весьма эффективна. Собственные значения всегда находятся в возрастающем порядке, и это важно, так как ленточные матрицы обычно возникают в конечноразностных приближениях дифференциальных уравнений, а там существенны только малые собственные значения. В целом кажется, что техника Рунтисхаузера предпочтительнее, хотя можно многое сказать в пользу обоих методов.

Несимметричные ленточные матрицы

66. Если мы используем обычную LR -технику без перестановок, то ленточная форма сохраняется. На самом деле ленточные матрицы даже не должны иметь одинаковое число элементов с каждой стороны диагонали. Однако численная устойчивость уже не обеспечивается, и мы не можем, вообще говоря, рекомендовать такое использование LR -алгоритма. Если используется LR -алгоритм с перестановками или QR -алгоритм, то, вообще говоря, ленточная форма *над диагональю* постепенно портится.

Особенно интересен случай трехдиагональной матрицы A . Сравним две последовательности трехдиагональных матриц A_s и $\tilde{A}_s$, полученных следующим образом. В первой начнем с A_1 и используем LR -разложение (без перестановок), в котором каждая L_s — единичная нижняя треугольная. Во второй начнем с DA_1D^{-1} , где D — некоторая диагональная матрица, и используем любой другой тип LR -разложения (без перестановок). Доказательство § 52 показывает, что на соответствующих стадиях мы имеем

$$\tilde{A}_s = D_s A_s D_s^{-1} \quad (66.1)$$

с некоторой диагональной матрицей D_s . Если

$$A_s = \begin{bmatrix} \alpha_1^{(s)} & \beta_2^{(s)} & & & \\ \gamma_2^{(s)} & \alpha_2^{(s)} & \beta_3^{(s)} & & \\ \dots & \dots & \dots & \dots & \\ & & \gamma_n^{(s)} & \alpha_n^{(s)} & \end{bmatrix}, \quad \tilde{A}_s = \begin{bmatrix} a_1^{(s)} & b_2^{(s)} & & & \\ c_2^{(s)} & a_2^{(s)} & b_3^{(s)} & & \\ \dots & \dots & \dots & \dots & \\ & & c_n^{(s)} & a_n^{(s)} & \end{bmatrix}, \quad (66.2)$$

то из (66.1) найдем

$$\alpha_i^{(s)} = a_i^{(s)}, \quad \beta_i^{(s)} \gamma_i^{(s)} = b_i^{(s)} c_i^{(s)}. \quad (66.3)$$

Следовательно, какое бы треугольное разложение мы ни использовали, диагональные элементы и произведения симметрично расположенных внедиагональных элементов совпадают, а это те величины, которые определяют проблему собственных значений для A_1 . Поэтому без потери общности мы можем выбрать D так, что

$$DA_1D^{-1} = \begin{bmatrix} \alpha_1 & 1 & & & \\ \beta_2 & \alpha_2 & 1 & & \\ \dots & \dots & \dots & \dots & \\ & & \beta_n & \alpha_n & \end{bmatrix}, \quad (66.4)$$

и затем использовать LR -разложение с единичной нижней треугольной L_s на каждой стадии. Если

$$\begin{bmatrix} \alpha_1^{(s)} & 1 & & & \\ \beta_2^{(s)} & \alpha_2^{(s)} & 1 & & \\ & \beta_3^{(s)} & \alpha_3^{(s)} & 1 & \\ \dots & \dots & \dots & \dots & \dots \\ & & & \beta_n^{(s)} & \alpha_n^{(s)} \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ l_2^{(s)} & 1 & & & \\ & l_3^{(s)} & 1 & & \\ \dots & \dots & \dots & \dots & \dots \\ & & & l_n^{(s)} & 1 \end{bmatrix} \begin{bmatrix} u_1^{(s)} & 1 & & & \\ & u_2^{(s)} & 1 & & \\ & & u_3^{(s)} & 1 & \\ \dots & \dots & \dots & \dots & \dots \\ & & & & u_n^{(s)} \end{bmatrix}, \quad (66.5)$$

то

$$\left. \begin{aligned} u_i^{(s)} &= \alpha_i^{(s)}, \quad l_i^{(s)} u_{i-1}^{(s)} = \beta_i^{(s)}, \quad l_i^{(s)} + u_i^{(s)} = \alpha_i^{(s)} \quad (i=2, \dots, n), \\ u_i^{(s)} + l_{i+1}^{(s)} &= \alpha_i^{(s+1)} \quad (i=1, \dots, n), \quad \beta_i^{(s+1)} = l_i^{(s)} u_i^{(s)} \quad (i=2, \dots, n) \end{aligned} \right\} \quad (66.6)$$

и, следовательно,

$$u_i^{(s)} + l_{i+1}^{(s)} = \alpha_i^{(s+1)} = l_i^{(s+1)} + u_i^{(s+1)}, \quad l_i^{(s)} u_i^{(s)} = \beta_i^{(s+1)} = l_i^{(s+1)} u_{i-1}^{(s+1)}. \quad (66.7)$$

Эти уравнения показывают, что после вычисления $u_i^{(1)}$, $l_i^{(1)}$ все остальные $u_i^{(s)}$ и $l_i^{(s)}$ могут быть определены без вычисления $\alpha_i^{(s)}$ и $\beta_i^{(s)}$. В самом деле, если мы рассматриваем $l_1^{(s)}$ как нуль при всех s , то имеем схему

$$\begin{array}{ccccccc} & & & & & & \\ 0 & & & & & & \\ & u_1^{(1)} & & & & & \\ & & l_2^{(1)} & & u_2^{(1)} & & \\ 0 & & & & & & \\ & u_1^{(2)} & & l_2^{(2)} & & u_2^{(2)} & l_3^{(1)} \\ & & & & & & \\ & u_1^{(3)} & & l_2^{(3)} & & u_2^{(3)} & l_3^{(2)} \\ & & & & & & \\ & & & & & & u_3^{(1)} \\ & & & & & & \\ & & & & & & u_3^{(2)} \\ & & & & & & \\ & & & & & & u_3^{(3)} \end{array} \quad (66.8)$$

в которой каждая наклонная линия с верхним индексом s может быть получена по выше лежащей наклонной линии, если использовать уравнения (66.7), причем связанные элементы размещаются по вершинам ромбов того типа, какой мы отметили. Читатель, знакомый с QD -алгоритмом, заметит, что $l_i^{(s)}$ и $u_i^{(s)}$ это $q_i^{(s)}$ и $c_i^{(s)}$ Рунтсхаузера. Алгоритм QD играет значительно большую роль в общей проблеме вычисления нулей мероморфных функций, и полное обсуждение его выходит за рамки этой книги.

С точки зрения вышеописанной проблемы QD -алгоритм кажется неподходящим, так как он соответствует исключению элементов без перестановок. Однако мы знаем, что если A_1 — положительно определенная, и мы все время используем разложение Холецкого, то численная устойчивость гарантируется. Если

$$\tilde{A}_s = \begin{bmatrix} a_1^{(s)} & b_2^{(s)} & & & \\ b_2^{(s)} & a_2^{(s)} & b_3^{(s)} & & \\ \dots & \dots & \dots & \dots & \dots \\ & & & b_n^{(s)} & a_n^{(s)} \end{bmatrix}, \quad (66.9)$$

то в этом случае наш анализ показывает, что $\alpha_i^{(s)}$ и $(b_i^{(s)})^2$ это точно те же величины, что и $\alpha_i^{(s)}$ и $\beta_i^{(s)}$, полученные при использовании формы (66.3). Кажется, что более удобно работать в терминах $\alpha_i^{(s)}$ и $\beta_i^{(s)}$. На самом деле это означает, что мы можем получить все преимущества LR Холецкого без вычисления квадратных корней. Вероятно, проще всего переходить прямо от $\alpha_i^{(s)}$, $\beta_i^{(s)}$ к α_i^{s+1} , β_i^{s+1} , и если мы используем на каждой стадии сдвиг k_s и не восстанавливающий метод, то уравнения будут иметь вид

$$\left. \begin{aligned} u_1^{(s)} &= \alpha_1^{(s)} - k_s, \\ l_i^{(s)} &= \beta_i^{(s)} / u_{i-1}^{(s)}, & u_{i-1}^{(s+1)} &= u_{i-1}^{(s)} + l_i^{(s)}, \\ u_i^{(s)} &= \alpha_i^{(s)} - k_s - l_i^{(s)}, & \beta_i^{(s+1)} &= l_i^{(s)} u_i^{(s)}, \\ & & \alpha_n^{(s+1)} &= u_n^{(s)}. \end{aligned} \right\} (i = 2, \dots, n), \quad (66.10)$$

При предположении, что полный сдвиг на всех стадиях меньше наименьшего собственного значения, все $\alpha_i^{(s)}$ и $\beta_i^{(s)}$ будут положительными и процесс будет устойчивым. На каждом шаге производится только $n - 1$ деление и $n - 1$ умножение.

Одновременное разложение и переумножение в QR-алгоритме

67. Ортега и Кайзер (1963) установили, что аналогичный процесс может быть использован для перехода от A_s к A_{s+1} при помощи QR-техники без использования квадратных корней. При этих преобразованиях требуется только симметрия, а не положительная определенность.

Опуская верхний индекс s , рассмотрим переход от A к $\bar{A}$ при помощи соотношений

$$A = Q_1 R_1, \quad R_1 Q_1 = \bar{A} \quad (67.1)$$

без использования сдвига. Приведение A к R достигается умножениями слева на $(n - 1)$ вращения в плоскостях $(i, i + 1)$, $(i = 1, \dots, n - 1)$. Если

$$A = \begin{bmatrix} a_1 & b_2 & & \\ b_2 & a_2 & b_3 & \\ & b_3 & a_3 & b_4 \\ \dots & \dots & \dots & \dots \\ & & b_n & a_n \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} \bar{a}_1 & \bar{b}_2 & & \\ \bar{b}_2 & \bar{a}_2 & \bar{b}_3 & \\ & \bar{b}_3 & \bar{a}_3 & \bar{b}_4 \\ \dots & \dots & \dots & \dots \\ & & \bar{b}_n & \bar{a}_n \end{bmatrix}, \quad (67.2)$$

$$R = \begin{bmatrix} r_1 & q_1 & t_1 & & \\ & r_2 & q_2 & t_2 & \\ & & r_3 & q_3 & t_3 \\ \dots & \dots & \dots & \dots & \dots \\ & & & & r_n \end{bmatrix},$$

а косинусы и синусы j -го вращения обозначены через c_j и s_j , то

$$s_j = b_{j+1} / (p_j^2 + b_{j+1}^2)^{1/2}, \quad c_j = p_j / (p_j^2 + b_{j+1}^2)^{1/2} \quad (j = 1, \dots, n - 1), \quad (67.3)$$

где

$$\left. \begin{aligned} p_1 &= a_1, & p_2 &= c_1 a_2 - s_1 b_2, \\ p_j &= c_{j-1} a_j - s_{j-1} c_{j-2} b_j & (j = 3, \dots, n) \end{aligned} \right\} \quad (67.4)$$

и

$$\left. \begin{aligned} r_j &= c_j p_j + s_j b_{j+1} & (j = 1, \dots, n-1), & & r_n &= p_n, \\ q_1 &= c_1 b_2 + s_1 a_2, & q_j &= c_j c_{j-1} b_{j+1} + s_j a_{j+1} & (j = 2, \dots, n-1), \\ t_j &= s_j b_{j+2} & (j = 1, \dots, n-2). \end{aligned} \right\} \quad (67.5)$$

С другой стороны, из переумножения имеем

$$\left. \begin{aligned} \bar{a}_1 &= c_1 r_1 + s_1 q_1, & \bar{a}_j &= c_{j-1} c_j r_j + s_j q_j & (j = 2, \dots, n-1), \\ \bar{a}_n &= c_{n-1} r_n, & \bar{b}_{j+1} &= s_j r_{j+1} & (j = 1, \dots, n-1). \end{aligned} \right\} \quad (67.6)$$

Определим величины

$$\gamma_1 = p_1, \quad \gamma_j = c_{j-1} p_j \quad (j = 2, \dots, n), \quad (67.7)$$

тогда из (67.6) и (67.5) найдем

$$\left. \begin{aligned} \bar{a}_j &= (1 + s_j^2) \gamma_j + s_j^2 a_{j+1}, & (j=1, \dots, n-1), & & \bar{a}_n &= \gamma_n, \\ \bar{b}_{j+1}^2 &= s_j^2 (p_{j+1}^2 + b_{j+2}^2) & (j = 1, \dots, n-2), \\ \bar{b}_n^2 &= s_{n-1}^2 p_n^2, \end{aligned} \right\} \quad (67.8)$$

а для γ_j и p_j^2 имеем

$$\left. \begin{aligned} \gamma_1 &= p_1 = a_1, & \gamma_j &= a_j - s_{j-1}^2 (a_j + \gamma_{j-1}) & (j = 2, \dots, n), \\ p_1^2 &= a_1^2, & p_j^2 &= \gamma_j^2 / c_{j-1}^2, & \text{если } c_{j-1} \neq 0, \\ & & p_j^2 &= c_{j-2}^2 b_j^2, & \text{если } c_{j-1} = 0. \end{aligned} \right\} \quad (67.9)$$

Окончательный алгоритм, следовательно, описывается уравнениями

$$u = 0, \quad c_0 = 1, \quad b_{n+1} = 0, \quad a_{n+1} = 0, \quad (67.10)$$

$$\left. \begin{aligned} \gamma_i &= a_i - u_{i-1}, \\ p_i^2 &= \gamma_i^2 / c_{i-1}^2, & \text{если } c_{i-1} \neq 0, \\ p_i^2 &= c_{i-2}^2 b_i^2, & \text{если } c_{i-1} = 0, \\ \bar{b}_i^2 &= s_{i-1}^2 (p_i^2 + b_{i+1}^2) & (i \neq 1), \\ s_i^2 &= b_{i+1}^2 / (p_i^2 + b_{i+1}^2), \\ c_i^2 &= p_i^2 / (p_i^2 + b_{i+1}^2), \\ u_i &= s_i^2 (\gamma_i + a_{i+1}), \\ \bar{a}_i &= \gamma_i + u_i. \end{aligned} \right\} \quad (i = 1, \dots, n). \quad (67.11)$$

Заметим, что мы вычисляем s_i^2 и c_i^2 , так как мы не можем использовать $(1 - s_i^2)$ для c_i^2 , если c_i мал. Каждая итерация требует выполнения $3n$ делений и $2n$ умножений, но ни одного извлечения квадратного корня. Величины $\bar{a}_i$ и $\bar{b}_i^2$ записываются на места a_i и b_i^2 .

Как и для общих симметричных ленточных матриц, оба алгоритма LR Холецкого и QR имеют свои преимущества, и выбор между ними затруднителен.

Уменьшение ширины ленточной матрицы

68. Техника §§ 59—64 весьма эффективна, если требуется найти только несколько собственных значений общей ленточной матрицы. Для вычисления большого числа собственных значений, может быть, лучше сначала привести матрицу к трехдиагональному виду при помощи предварительного преобразования. Это можно сделать при помощи методов Гивенса или Хаусхолдера, но ширина матрицы на промежуточных стадиях приведения увеличивается.

Рутисхаузер (1963) описал два интересных приведения, использующих соответственно плоские вращения и матрицы отражения, в которых ленточная симметричная форма сохраняется на всех стадиях. Мы дадим только приведение, основанное на плоских вращениях, и в нашем описании используем только элементы верхнего треугольника.

Предположим, что ширина ленты равна $(2m + 1)$. Сначала аннулируем элемент $(1, m + 1)$, оставляя ленту шириной $(2m + 1)$ в строках и столбцах от 2 до n . Все вместе требует $p = [(n - 1)/m]$ вращений (здесь $[x]$ означает целую часть x), так как, аннулируя $(1, m + 1)$ элемент, вращение вводит отличный от нуля элемент как раз вне нижней полосы снизу. Он в свою очередь должен быть аннулирован, и порядок вычислений может быть представлен таблицей.

Аннулируемый элемент	Плоскость вращения	Новый введенный элемент
$1, m + 1$	$m, m + 1$	$m, 2m + 1$
$m, 2m + 1$	$2m, 2m + 1$	$2m, 3m + 1$
$2m, 3m + 1$	$3m, 3m + 1$	$3m, 4m + 1$
$\vdots$	$\vdots$	$\vdots$
$(p - 1)m, pm + 1$	$pm, pm + 1$	нет

После этого первые строка и столбец матрицы будут ширины $2m - 1$, а оставшаяся часть — ширины $2m + 1$. Следующий шаг — аннулирование элемента $(2, m + 2)$. Это выполняется тем же самым образом и требует $[(n - 2)/m]$ вращений. Продолжая процесс, мы получаем, что вся матрица постепенно приводится к матрице шириной $2m - 1$. Абсолютно аналогичный процесс приводит ее к матрице шириной $2m - 3$, и следовательно, матрица непременно станет трехдиагональной.

Число вращений, необходимое для приведения матрицы ширины $2m + 1$ к матрице ширины $2m - 1$, равно $[(n - 1)/m] + [(n - 2)/m] + \dots + [(n - 3)/m] + \dots$, что приблизительно равно $4(m + 1)n^2/2m$, а число умножений приблизительно равно $4(m + 1)n^2/m$. Процесс наиболее часто используется, когда $m = 2$ или 3.

Дополнительные замечания

Предположение о том, что методы, аналогичные методу Якоби, могли бы быть использованы для приведения матриц общего вида к треугольному виду, делалось фон Нейманом и Козеем (1958); Гринштадт (1955) и Лоткин (1956) описали методы такого типа. На конференции по теории матриц в Геттингбурге в 1961 г. Гринштадт дал обзор методов этой группы, предложенных к тому времени, и сделал вывод, что еще не развито ни одной удовлетворительной процедуры. Эберлейн (1962) описала модификацию метода Якоби, основанную на замечании, что для любой матрицы A существует преобразование подобия $B = P^{-1}AP$, которое сколь угодно близко к нормальной матрице. В алгоритме Эберлейн P строится в виде произведения последовательности матриц, которые являются обобщениями плоских вращений, но уже не являются унитарными. Итерации продолжаются до тех пор, пока B не

станет с рабочей точностью нормальной, и основная черта метода состоит в том, что, вообще говоря, предельная матрица B является прямой суммой некоторого числа матриц 1-го и 2-го порядков, так что можно определить все собственные значения.

Эберлейн также рассматривала приведение матриц общего вида к нормальному виду, используя комбинации плоских вращений и диагональных преобразований подобия, и подобные идеи независимо развивались Рутисхаузером. Обычно ведущей стратегией на каждой стадии является уменьшение отклонения Хенричи (гл. 3, § 50). Развитие этой линии может привести к методам, которые превзойдут те методы, которые я описал, и, может быть, это будет наиболее перспективной линией развития.

Методы этого класса не покрываются никаким из общих анализов ошибок, которые я дал. Однако можно ожидать, что они будут устойчивыми, так как последовательные приведенные матрицы стремятся к нормальной матрице, проблема собственных значений для которой исключительно хорошо обусловлена.

Первое упоминание LR -алгоритма было сделано Рутисхаузером в 1955 г. С тех пор он постоянно развивал и расширял свою теорию в серии работ. Я попытался здесь обобщить наиболее позднюю информацию, полученную из работ Рутисхаузера и моего собственного опыта работы в Национальной физической лаборатории. Я включил доказательства сходимости, данные Рутисхаузером в его работах, так как они являются типичными. LR -алгоритм со сдвигом и перестановками в применении к матрицам Хессенберга был впервые использован автором в 1959 г. в основном для матриц с комплексными элементами.

Работы Френсиса по QR -алгоритму помечены 1959 г., но не опубликованы вплоть до 1961 г. Одновременно и независимо тот же алгоритм был открыт Кублановской (1961). В свете нашего параллельного изучения ортогональных и неортогональных преобразований он является естественной аналогией LR -алгоритма, но работа Френсиса содержит значительно больше, чем простое описание QR -алгоритма. Обе техники — для комбинирования комплексно сопряженных сдвигов и описанная в § 38 техника, имеющая дело с последовательными малыми поддиагональными элементами, — являются важным вкладом в эффективность QR -алгоритма.

Доказательство сходимости QR -алгоритма, данное в §§ 29—32, было найдено автором во время написания настоящей главы. Доказательства, основанные на более трудоемкой теории определителей, даны Кублановской (1961) и Хаусхолдером (1964).

Уменьшение ширины ленточной матрицы в симметричном случае, как при помощи плоских вращений, так и при помощи матриц отражения, было описано Рутисхаузером (1963) в работе, которая включала также много других интересных преобразований, основанных на элементарных ортогональных матрицах.

ИТЕРАЦИОННЫЕ МЕТОДЫ

Введение

1. В этой последней главе мы будем иметь дело с тем, что обычно называют итерационными методами решения проблемы собственных значений. Это название несколько неточно, так как вся техника получения собственных значений по существу итеративна. Отличительной чертой методов этой главы является то, что собственные значения находятся лишь после определения собственного вектора или, возможно, нескольких собственных векторов. Читатель мог заметить, что раньше мы мало говорили о вычислении собственных векторов неэрмитовых матриц; это потому, что наиболее устойчивыми способами вычисления собственных векторов являются способы «итерационного» типа. Интересно заметить, что устойчивые методы для вычисления собственных значений не обязательно ведут к устойчивым методам для собственных векторов. Мы уже обращали на это внимание в связи с вычислением собственных значений трехдиагональных матриц (гл. 5, §§ 48—52).

Методы, описанные в этой главе, определены значительно менее точно, чем методы предыдущих глав. Хотя большей частью они идейно очень просты, им трудно придать вид автоматической процедуры. Это, в частности, верно для тех способов, которые предназначены выявлять присутствие нелинейных элементарных делителей и определять соответствующие инвариантные подпространства. На практике значительно легче иметь дело с нелинейными элементарными делителями, если заранее известно, что они присутствуют.

Исключением из общего правила является метод обратной итерации, описанный в §§ 47—60. Это наиболее употребительный метод для вычисления собственных векторов.

Степенной метод

2. Ограничимся рассмотрением, за исключением специально оговоренных случаев, матриц, имеющих линейные элементарные делители. Для любой такой матрицы A имеем

$$A = X \operatorname{diag}(\lambda_i) X^{-1} = X \operatorname{diag}(\lambda_i) Y^T = \sum_1^n \lambda_i x_i y_i^T, \quad (2.1)$$

где столбцы x_i матрицы X и строки y_i^T матрицы Y^T являются правыми и левыми собственными векторами A , нормированными так, что

$$y_i^T x_i = 1. \quad (2.2)$$

Следовательно,

$$A^s = X \operatorname{diag}(\lambda_i^s) Y^T = \sum_1^n \lambda_i^s x_i y_i^T, \quad (2.3)$$

и если

$$|\lambda_1| = |\lambda_2| = \dots = |\lambda_r| > |\lambda_{r+1}| \geq \dots \geq |\lambda_n|, \quad (2.4)$$

то главный член в правой части (2.3) есть $\sum_1^r \lambda_i^s x_i y_i^T$. Это основной результат, на котором основаны все методы этой главы. (Мы уже видели, что это является основанием *LR*- и *QR*- алгоритмов гл. 8.)

Мы будем называть собственные значения $\lambda_1, \dots, \lambda_r$ *доминирующими собственными значениями*, а соответствующие собственные векторы *доминирующими собственными векторами*. Наиболее часто $r = 1$, и в этом случае главный член в A^s есть $\lambda_1^s x_1 y_1^T$.

Большинство практических способов основано на так называемом *степенном методе* вместе с приемами, позволяющими увеличить доминирование собственного значения, наибольшего по модулю. Последние используют тот факт, что если $p(A)$ и $q(A)$ — полиномы от A , то

$$p(A) \{q(A)\}^{-1} = X \operatorname{diag} \{p(\lambda_i)/q(\lambda_i)\} Y^T \quad (2.5)$$

при условии, что $q(A)$ — неособенная. Если мы обозначим

$$\mu_i = p(\lambda_i)/q(\lambda_i), \quad (2.6)$$

то целью является такой выбор полиномов $p(A)$ и $q(A)$, чтобы один из $|\mu_i|$ был много больше других.

Прямые итерации одного вектора

3. Простейшее применение степенного метода таково. Пусть u_0 произвольный вектор, и пусть последовательности векторов v_s и u_s определяются уравнениями

$$v_{s+1} = A u_s, \quad u_{s+1} = v_{s+1} / \max(v_{s+1}); \quad (3.1)$$

здесь и в дальнейшем мы используем обозначение $\max(x)$ для максимального по модулю элемента вектора x . Очевидно, мы имеем

$$u_s = A^s u_0 / \max(A^s u_0), \quad (3.2)$$

и если мы положим, что

$$u_0 = \sum_1^n \alpha_i x_i, \quad (3.3)$$

то, с точностью до нормирующего множителя, u_s имеет вид

$$\sum_1^n \alpha_i \lambda_i^s x_i = \lambda_1^s \left[\alpha_1 x_1 + \sum_2^n \alpha_i (\lambda_i/\lambda_1)^s x_i \right]. \quad (3.4)$$

Если $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$, то, предполагая $\alpha_1 \neq 0$, получим

$$u_s \rightarrow x_1 / \max(x_1) \quad \text{и} \quad \max(v_s) \rightarrow \lambda_1. \quad (3.5)$$

Следовательно, этот процесс дает одновременно доминирующее собственное значение и соответствующий собственный вектор. Если $|\lambda_1/\lambda_2|$ близко к единице, то сходимость будет очень медленная.

Если существует несколько линейно независимых собственных векторов, соответствующих доминирующему собственному значению, то это

не влияет на сходимость. Действительно, если

$$\lambda_1 = \lambda_2 = \dots = \lambda_r \quad \text{и} \quad |\lambda_1| > |\lambda_{r+1}| \geq \dots \geq |\lambda_n|, \quad (3.6)$$

то

$$A^s u_0 = \lambda_1^s \left[\sum_1^r \alpha_i x_i + \sum_{r+1}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^s x_i \right] \sim \lambda_1^s \sum_1^r \alpha_i x_i. \quad (3.7)$$

Итерации поэтому сходятся к некоторому вектору, лежащему в подпространстве, натянутом на собственные векторы $x_1, \dots, x_r$, причем предел зависит от начального вектора u_0 .

Сдвиг начала

4. Простейшие полиномы по A имеют вид $(A - pI)$. Соответствующие собственные значения равны $\lambda_i - p$, и если p выбрать соответствующим образом, сходимость к собственному вектору может быть ускорена. Мы рассмотрим сначала случай, когда A и все λ_i вещественны. Ясно, что если $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_{n-1} > \lambda_n$, то как бы мы ни выбирали p , доминирующим собственным значением будет либо $\lambda_1 - p$, либо $\lambda_n - p$. Для сходимости к x_1 наилучшим значением p будет $\frac{1}{2}(\lambda_2 + \lambda_n)$, и быстрота сходимости определится скоростью, с которой

$$\{(\lambda_2 - \lambda_n)/(2\lambda_1 - \lambda_2 - \lambda_n)\}^s$$

стремится к нулю. Аналогично оптимальным выбором для сходимости к x_n будет $p = \frac{1}{2}(\lambda_1 + \lambda_{n-1})$, и быстрота сходимости определяется величиной $\{(\lambda_1 - \lambda_{n-1})/(\lambda_1 + \lambda_{n-1} - 2\lambda_n)\}^s$. Проанализируем сначала эффект от употребления специальных значений p , оставив проблему определения таких значений на будущее.

Для некоторых расположений собственных значений этот простой прием может быть весьма эффективным. Например, для матрицы шестого порядка с $\lambda_i = 21 - i$ итерации без сдвига дают сходимость к x_1 со скоростью $\{19/20\}^s$. Если мы возьмем $p = 17$, то собственные значения станут $3, \pm 2, \pm 1, 0$, и мы все еще будем иметь сходимость к x_1 , но со скоростью $(2/3)^s$. Для получения x_1 с определенной точностью требуется приблизительно в восемь раз больше итераций, чем при ускоренном варианте. Аналогично, если мы возьмем $p = 18$, собственные значения станут $\pm 2, \pm 1, 0, 0, -3$, и мы получим сходимость к x_6 со скоростью $(2/3)^s$.

Довольно часто бывают такие неблагоприятные расположения собственных значений, для которых максимально достижимая таким приемом скорость сходимости неудовлетворительна. Пусть, например,

$$\lambda_i = 1/i \quad (i = 1, \dots, 20) \quad (4.1)$$

и мы хотим выбрать p так, чтобы получить x_{20} . Оптимальный выбор p это $1/2(1 + 1/19)$, и скорость сходимости определяется величиной $(180/181)^s$. Для уменьшения относительной ошибки в векторе в e раз требуется приблизительно 181 итерация. Весьма обычны расположения собственных значений, для которых $|\lambda_{n-1} - \lambda_n| \ll |\lambda_1 - \lambda_n|$, при этом сходимость к x_n медленная даже с оптимальным выбором p .

Влияние ошибок округления

5. Прежде чем рассматривать более сложные способы ускорения сходимости, важно оценить влияние ошибок округления. Можно было бы подумать, что если мы работаем прямо с матрицей A , то достижимая точность будет высока, даже если матрица плохо обусловлена в смысле проблемы собственных векторов; но это не так. Если мы учтем влияние ошибок округления в процессе § 3, то, обозначая $\max (v_{s+1}) = k_{s+1}$, получим

$$k_{s+1}u_{s+1} = Au_s + f_s, \quad (5.1)$$

где, вообще говоря, f_s это «малый» ненулевой вектор. Поэтому весьма возможно, что u_{s+1} станет равным u_s , хотя еще будет далек от точного собственного вектора.

Рассмотрим, например, матрицу

$$A = \begin{bmatrix} 0,9901 & 10^{-3}(0,2000) \\ 10^{-3}(-0,1000) & 0,9904 \end{bmatrix}, \quad (5.2)$$

для которой

$$\left. \begin{aligned} \lambda_1 &= 0,9903, & x_1^T &= (1, 1), \\ \lambda_2 &= 0,9902, & x_2^T &= (1, 0,5). \end{aligned} \right\} \quad (5.3)$$

Если мы возьмем $u_s^T = (1, 0,9)$, то точные вычисления дадут

$$u_{s+1}^T = (0,99028, 0,89126), \quad u_{s+1}^T = (1,0, 0,900008 \dots), \quad (5.4)$$

$$k_{s+1} = 0,99028.$$

Разность между компонентами u_{s+1} и u_s меньше чем $5 \cdot 10^{-5}$, и если мы работаем с четырьмя десятичными знаками, то ясно, что вычисленный вектор u_{s+1} будет в точности равен u_s , и процесс станет стационарным. Можно проверить, что такое поведение характерно для широкого класса векторов u_s .

6. Если наибольшей по модулю компонентой вектора u_s является i -я, то (5.1) может быть записано в виде

$$k_{s+1}u_{s+1} = (A + f_s e_i^T) u_s, \quad (6.1)$$

и следовательно, мы по существу выполнили точную итерацию с $A + f_s e_i^T$, а не с A . Следовательно, получим собственный вектор матрицы $A + f_s e_i^T$. (Существует, конечно, много других возмущений F матрицы A таких, что $(A + F) u_s = k_{s+1} u_{s+1}$.) В соответствии с данным способом вычисления Au_s в действительности будет получаться область неопределенности, связанная с каждым собственным вектором, и процесс получения точного собственного вектора будет, вообще говоря, прекращаться, когда будет достигнут вектор, лежащий в этой области.

Если, например, Au_s вычислен с использованием стандартной арифметики с плавающей запятой, то мы имеем из гл. 3, § 6 (v)

$$fl(Au_s) = (A + F_s)u_s, \quad |F_s| < 2^{-t_1} |A| \operatorname{diag}(n+1-i). \quad (6.2)$$

Мы не можем гарантировать, что приближение будет продолжаться после того, как мы получим вектор u_s , который является точным собственным вектором некоторой матрицы $A + G$, где

$$|G| < 2^{-t_1} |A| \operatorname{diag}(n+1-i), \quad (6.3)$$

хотя обычно ошибка в векторе u_s предельной точности значительно меньше, чем максимум, который может вызвать возмущение, удовлетворяющее (6.2). Заметим, что граница для каждой компоненты $|f_{ij}|$ в (6.2) равна $2^{-t_1}(n+1-j)|a_{ij}|$, и так как она прямо пропорциональна $|a_{ij}|$, то ясно, что так же, как и в гл. 7, § 15, точность, достижимая при вычислении доминирующего собственного значения итерациями с $D^{-1}AD$ (где D — любая диагональная матрица), такая же, как и при итерировании с A .

В примере § 5 k_{s+1} предельной точности является хорошим приближением к λ_1 , несмотря на неточность вектора, из которого оно было определено. Это вызвано тем, что матрица имеет хорошо обусловленные *собственные значения*, но плохо обусловленные *собственные векторы*, так как собственные значения близки. Если λ_1 — плохо обусловленное собственное значение, т. е. если s_1 мало (гл. 2, § 34), то предельная точность, достижимая в собственном значении, также мала.

Рассмотрим, например, матрицу

$$\begin{bmatrix} 1 & 1 & 0 \\ -1 + 10^{-8} & 3 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad (6.4)$$

собственные значения которой 2 ± 10^{-4} и 1. Если мы возьмем $u_s^T = (1, 1, 1)$, то

$$(Au_s)^T = (2, 2 + 10^{-8}, 2), \quad (6.5)$$

так что при использовании арифметики с плавающей запятой и восьми-значной мантиссой, процесс стационарен и дает в качестве собственного значения $\lambda = 2$.

7. Мы видим, что влияние ошибок округления, сделанных в итерационном процессе, не имеет существенно другой природы, чем влияние соответствующих ошибок, сделанных при последовательных преобразованиях подобия. Тем не менее мы знаем из собственного опыта, что если, например, мы итерируем для получения доминирующего собственного вектора, используя *стандартную* арифметику с плавающей запятой, то вводимая ошибка обычно больше, чем та, которая соответствует ошибке, сделанной при приведении к форме Хессенберга с использованием матриц отражения с *плавающей запятой и накоплением*.

Что касается матрицы (5.2), то можно точно получить собственные векторы, используя подходящий сдвиг. Имеем, например,

$$A - 0,99I = \begin{bmatrix} 10^{-3}(0,1000) & 10^{-3}(0,2000) \\ 10^{-3}(0,1000) & 10^{-3}(0,4000) \end{bmatrix}, \quad (7.1)$$

и эта матрица имеет собственные значения $10^{-3}(0,3)$ и $10^{-3}(0,2)$. Это весьма частный случай. Вообще говоря, если $|\lambda_1 - \lambda_2| \ll |\lambda_1 - \lambda_n|$, то трудности при точном вычислении x_1 и x_2 являются фундаментальными, и их не обойти таким простым приемом.

8. Существует другой вопрос, в котором ошибки округления имеют глубокое влияние. Предположим, например, что мы имеем матрицу A третьего порядка с собственными значениями порядка 1,0, 0,9 и 10^{-4} . Если мы итерируем прямо с A , то при точных вычислениях компонента по x_3 уменьшится при s итерациях, получив множитель 10^{-4s} . Однако

при округлении каждого элемента u_s до t двоичных знаков будет вводиться компонента по x_3 , которая будет, вообще говоря, порядка $2^{-t}x_3$. На самом деле обычно не существует t -значных приближений к x_1 , которые содержат компоненты по x_3 существенно меньшие, чем это. Важность этого замечания станет понятной из следующего параграфа.

Изменение p

9. Для получения наибольших преимуществ при последовательных итерациях величину p можно менять. Если мы обозначим текущий вектор $\sum \alpha_i x_i$, то после одной итерации он станет с точностью до нормирующего множителя $\sum \alpha_i (\lambda_i - p) x_i$. Если $p = \lambda_k$, то компонента по x_k полностью исчезнет, а если p близко к λ_k , то значительно уменьшится относительная величина компоненты по x_k . Предположим, что мы достигли стадии, на которой

$$u_s = \alpha_1 x_1 + \alpha_2 x_2 + \sum_{i=3}^n \varepsilon_i x_i, \quad |\varepsilon_i| \ll |\alpha_1|, |\alpha_2|, \quad (9.1)$$

и мы имеем некоторую оценку λ'_2 величины λ_2 . Взяв $p = \lambda'_2$, получим

$$(A - pI)u_s = \alpha_1 (\lambda_1 - \lambda'_2) x_1 + \alpha_2 (\lambda_2 - \lambda'_2) x_2 + \sum_3^n \varepsilon_i (\lambda_i - \lambda'_2) x_i. \quad (9.2)$$

Если $|\lambda_2 - \lambda'_2| \ll |\lambda_1 - \lambda'_2|$, то компонента по x_2 относительно сильнее уменьшилась, чем компонента по x_1 , хотя если $|\lambda_i - \lambda'_2| > |\lambda_1 - \lambda'_2|$ для некоторого $i > 2$, то соответствующая компонента увеличивается по сравнению с компонентой по x_1 . Если ε_i малы, то мы можем применить несколько итераций с $p = \lambda'_2$, но мы не можем продолжать это бесконечно, так как иначе начнет доминировать компонента по x_i .

Для иллюстрации предположим, что матрица четвертого порядка имеет собственные значения 1,0, 0,9, 0,2, 0,1. Если мы итерируем без сдвига, то компоненты по x_3 и по x_4 будут быстро уменьшаться, но ошибки округления не дадут им исчезнуть. Предположим, что, работая с 10 десятичными знаками, мы достигли u_s вида

$$x_1 + 10^{-2}x_2 + 10^{-10}x_3 + 10^{-10}x_4 \quad (9.3)$$

и что на этом шаге имеем для λ_2 оценку 0,901. Если мы применим r итераций с $p = 0,901$, то получим

$$(A - pI)^r u_s = (0,099)^r x_1 + 10^{-2}(-0,001)^r x_2 + \\ + (-0,701)^r 10^{-10}x_3 + (0,801)^r 10^{-10}x_4. \quad (9.4)$$

Мешающая компонента по x_2 быстро уменьшается, но мы должны помнить, что при нормировании u_i ошибки округления будут всегда увеличивать компоненту по x_2 до величины 10^{-10} . При $r = 4$ видим, что u_{s+4} приблизительно пропорционален

$$10^{-4}(0,96)x_1 + 10^{-14}x_2 + 10^{-10}(0,24)x_3 + 10^{-10}(0,41)x_4,$$

т. е.

$$x_1 + 10^{-10}(1,04)x_2 + 10^{-6}(0,25)x_3 + 10^{-6}(0,43)x_4. \quad (9.5)$$

Дальнейшие итерации с $p = 0,901$ не будут уменьшать компоненты по x_2 , а компоненты по x_3 и по x_4 будут расти. Если мы теперь вернемся к $p = 0$, то компоненты по x_3 и по x_4 будут снова быстро уменьшаться, и

после пяти или шести итераций получим

$$x_1 + 10^{-10}x_2 + 10^{-10}x_3 + 10^{-10}x_4. \quad (9.6)$$

Нужно было бы приблизительно 200 итераций, для того чтобы получить такой результат, исходя из (9.3) и все время используя $p = 0$. Если λ_1 и λ_2 еще более близки, то такой способ дает даже большее преимущество, но для его применения надо знать лучшее приближение для λ_2 .

Специальный выбор p

10. На вычислительных машинах Национальной физической лаборатории была разработана специальная *ad hoc* версия такой техники, в которой выбор величины p осуществлялся оператором. Эта величина p задавалась на ручном пульте и считывалась машиной при начале каждой итерации, а величина текущей оценки $[\max(v_s) + p]$ для собственного значения в конце каждой итерации выдавалась на пульте и последовательные векторы u_s получались в двоичной системе на экране осциллографа.

Употребление таких программ, вероятно, лучше всего проиллюстрировать обсуждением одного специального приложения. Рассмотрим определение доминирующего собственного значения и собственного вектора матрицы, о которой известно, что ее собственные значения вещественны и положительны. Итерации начинаются при значении $p = 0$. Если сходимость достаточно быстрая, то нет надобности использовать какое-либо другое значение p , но если нет, то можно оценить скорость сходимости u_s . Так как программа использует арифметику с фиксированной запятой, это можно легко сделать, заметив, как много нужно итераций для того, чтобы значение каждого двоичного знака u_s стало стационарным. Предположим, что требуется k итераций на знак. Тогда мы можем считать, что

$$(\lambda_2/\lambda_1)^k \approx \frac{1}{2}, \quad (10.1)$$

и, взяв текущее значение $[\max(v_s) + p]$ в качестве λ_1 , получим

$$\lambda_2 = \{\max(v_s) + p\} (1/2)^{1/k}. \quad (10.2)$$

Это приближение для λ_2 подается на пульт. Обычно векторы u_i тогда показывают значительно более высокую скорость сходимости на нескольких следующих итерациях, но постепенно компоненты по $x_n, x_{n-1}, \dots$ приобретают значение, и знаки u_i , которые уже были найдены, начинают нарушаться.

На практике обычно так продолжают до тех пор, пока почти все знаки u_i не сменятся; на этой стадии компонента по x_n , вероятно, будет несколько больше, чем компонента по x_1 . Затем возвращаемся к значению $p = 0$, и компоненты по x_n и x_{n-1} уменьшаются снова. Если λ_1 и λ_2 весьма близки, то иногда необходимо повторять этот процесс несколько раз для получения x_1 с требуемой точностью.

Эта техника оказалась исключительно успешной для матриц высокого порядка, если только λ_1 и λ_2 не патологически близки. На практике рациональнее использовать величину p несколько меньшую, чем в (10.2), так как надежнее иметь p меньшим λ_2 , чем большим.

Процесс ускорения Эйткена

11. Эйткен (1937) описал процесс ускорения сходимости векторов, который иногда очень полезен. Предположим, что u_s, u_{s+1}, u_{s+2} — три последовательные итерации и что мы достигли стадии, на которой u_s по существу уже равен $x_1 + \varepsilon x_2$, где ε мало. Если x_1 и x_2 нормированы так, что $\max(x_i) = 1$, то для достаточно малых ε наибольший элемент u_s будет на том же месте, что и у x_1 . Предположим, что элемент x_2 на этом месте равен k , так что $|k| \leq 1$. Тогда для u_s, u_{s+1}, u_{s+2} имеем

$$(x_1 + \varepsilon x_2)/(1 + \varepsilon k), \quad (\lambda_1 x_1 + \varepsilon \lambda_2 x_2)/(\lambda_1 + \varepsilon k \lambda_2), \quad (\lambda_1^2 x_1 + \varepsilon \lambda_2^2 x_2)/(\lambda_1^2 + \varepsilon k \lambda_2^2). \quad (11.1)$$

Рассмотрим теперь вектор a с компонентами:

$$\begin{aligned} a_i &= [u_i^{(s)} u_i^{(s+2)} - (u_i^{(s+1)})^2] / [u_i^{(s)} - 2u_i^{(s+1)} + u_i^{(s+2)}] = \\ &= u_i^{(s)} - [u_i^{(s)} - u_i^{(s+1)}]^2 / [u_i^{(s)} - 2u_i^{(s+1)} + u_i^{(s+2)}], \end{aligned} \quad (11.2)$$

где $u_i^{(s)}$ означает i -ю компоненту u_s . Подставляя (11.1) и обозначая $\lambda_2/\lambda_1 = r$, после упрощений получим

$$a_i = \frac{\varepsilon(1-r)^2(x_i^{(2)} - kx_i^{(1)})(x_i^{(1)} - \varepsilon^2 r^2 k x_i^{(2)})}{\varepsilon(1-r)^2(x_i^{(2)} - kx_i^{(1)})(1 - \varepsilon^2 k^2 r^2)} = \quad (11.3)$$

$$= [x_i^{(1)} - \varepsilon^2 r^2 k x_i^{(2)}] / (1 - \varepsilon^2 k^2 r^2), \quad (11.4)$$

что дает

$$a = x_1 + O(\varepsilon^2). \quad (11.5)$$

Можно было бы подумать, что так как числитель и знаменатель a_i содержат множитель ε , то процесс серьезно зависит от ошибок округления; но это не так. Если u_s, u_{s+1}, u_{s+2} имеют ошибки, которые обозначим $2^{-t}\alpha, 2^{-t}\beta, 2^{-t}\gamma$, то, пренебрегая множителем $2^{-t}\varepsilon$, получим вместо (11.3)

$$a_i = \frac{\varepsilon(1-r^2)(x_i^{(2)} - kx_i^{(1)})(x_i^{(1)} - \varepsilon^2 r^2 k x_i^{(2)}) + x_i^{(1)} 2^{-t}(\alpha_i + \gamma_i - 2\beta_i)}{\varepsilon(1-r^2)(x_i^{(2)} - kx_i^{(1)})(1 - \varepsilon^2 k^2 r^2) + 2^{-t}(\alpha_i + \gamma_i - 2\beta_i)}. \quad (11.6)$$

Мы можем написать это в виде

$$a_i = \frac{px_i^{(1)} - \varepsilon^2 q x_i^{(2)}}{p - \varepsilon^2 q}, \quad (11.7)$$

откуда видно, что ошибки округлений сравнительно незначительны.

Следовательно, если мы достигли стадии, в которой влияние всех собственных векторов, кроме x_1, x_2 уже исключено, то вектор, который дается процессом Эйткена, более близок к x_1 , чем векторы, из которых он получен. Однако если λ_3 и λ_2 плохо разделены, то может потребоваться много итераций, прежде чем компонента по x_3 будет мала по сравнению с компонентой по x_2 .

Неожиданно трудно составить автоматическую программу, которая бы эффективно использовала процесс Эйткена. В *ad hoc* программе Национальной физической лаборатории использование процесса Эйткена контролировалось оператором. С помощью пульта он мог потребовать применения процесса в конце текущей итерации и, наблюдая за поведением итерированного вектора, мог оценить, был ли полезен процесс. Если

выигрыша не получалось, то использование процесса Эйткена для определения текущего вектора обычно блокировалось. Несмотря на трудности в осуществлении автоматического процесса, использование метода Эйткена имеет весьма большое значение.

Комплексно сопряженные собственные значения

12. Если доминирующие собственные значения действительной матрицы суть комплексно сопряженная пара λ_1 и $\bar{\lambda}_1$, то итерированные векторы не сходятся. В самом деле, если x_1 и $\bar{x}_1$ суть соответствующие собственные векторы, то произвольный вещественный вектор представим в виде

$$u_0 = \alpha_1 x_1 + \bar{\alpha}_1 \bar{x}_1 + \sum_3^n \alpha_i x_i. \quad (12.1)$$

Следовательно, имеем

$$A^s u_0 = r_1^s [\rho_1 e^{i(\alpha+s\theta)} x_1 + \rho_1 e^{-i(\alpha+s\theta)} \bar{x}_1 + \sum_3^n \alpha_i (\lambda_i/r_1)^s x_i], \quad (12.2)$$

где

$$\lambda_1 = r e^{i\theta}, \quad \alpha_1 = \rho_1 e^{i\alpha}. \quad (12.3)$$

Компоненты по $x_3, \dots, x_n$ обязательно исчезнут, но если положить

$$v_{s+1} = A u_s, \quad \max(v_{s+1}) = k_{s+1}, \quad u_{s+1} = v_{s+1}/k_{s+1}, \quad (12.4)$$

то ясно из (12.2), что ни k_{s+1} , ни u_{s+1} не стремятся к пределу. Если мы обозначим j -ю компоненту x_1 через $\xi_j e^{i\varphi_j}$, равенство (12.2) даст

$$(A^s u_0)_j = 2\rho_1 r_1^s \xi_j \cos(\alpha + \varphi_j + s\theta), \quad (12.5)$$

и, следовательно, компоненты u_s осциллируют.

Если λ_1 и $\bar{\lambda}_1$ суть корни уравнения $\lambda^2 - p\lambda - q = 0$, то

$$(A^{s+2} - pA^{s+1} - qA^s)u_0 \rightarrow 0 \quad (s \rightarrow \infty), \quad (12.6)$$

или

$$k_{s+1}k_{s+2}u_{s+2} - pk_{s+1}u_{s+1} - qu_s \rightarrow 0, \quad (12.7)$$

и, следовательно, обязательно любые три последовательные итерации будут линейно зависимыми. С помощью метода наименьших квадратов мы можем определять последовательные приближения p_s и q_s для p и q .

Имеем

$$k_{s+1}k_{s+2} \begin{bmatrix} u_{s+1}^T & u_{s+2} \\ u_s^T & u_{s+1} \end{bmatrix} = \begin{bmatrix} u_{s+1}^T u_{s+1} & u_{s+1}^T u_s \\ u_s^T u_{s+1} & u_s^T u_s \end{bmatrix} \begin{bmatrix} p_s & k_{s+1} \\ q_s \end{bmatrix}. \quad (12.8)$$

Если p_s и q_s стремятся к пределам p и q , то λ_1 и $\bar{\lambda}_1$ можно сосчитать из соотношений

$$\operatorname{Re}(\lambda_1) = \frac{1}{2}p, \quad \operatorname{Im}(\lambda_1) = \frac{1}{2}(p^2 + 4q)^{1/2}. \quad (12.9)$$

13. Так как λ_1 и $\bar{\lambda}_1$ определяются из алгебраического уравнения, они плохо определяются по коэффициентам p и q , если корни близки, т. е. если $\operatorname{Im}(\lambda_1)$ мала. Можно было бы подумать, что потеря точности в собственных значениях соответствует плохой обусловленности исходной задачи, но это не обязательно верно.

Рассмотрим, например, матрицу

$$\begin{bmatrix} 0,4812 & 0,0023 \\ -0,0024 & 0,4810 \end{bmatrix}. \quad (13.1)$$

Собственные значения ее равны $0,4811 \pm 0,0023i$, и можно легко проверить, что они сравнительно нечувствительны к малым возмущениям элементов A . Однако характеристическое уравнение

$$\lambda^2 - 0,9622\lambda + 0,23146272 = 0$$

имеет корни, весьма чувствительные к *независимым* изменениям коэффициентов. Следовательно, даже сравнительно точно определенные по методу § 12 p и q будут давать плохое значение для мнимой части собственного значения. В этом случае уравнения (12.8) заведомо плохо обусловлены.

Это хорошо иллюстрируется матрицей (13.1). Так как она второго порядка, то за u_s мы можем взять любой вектор. Если мы возьмем $u_s^T = (1, 1)$, то получим, работая с 4 десятичными знаками

$$(Au_s)^T = (0,4835, 0,4786), \quad (A^2u_s)^T = (0,2338, 0,2290). \quad (13.2)$$

Так как здесь лишь две компоненты, p и q определяются единственным образом, и мы имеем

$$0,2338 = 0,4835p + q,$$

$$0,2290 = 0,4786p + q,$$

что дает

$$p = 0,9796, \quad q = -0,2398. \quad (13.3)$$

Точные вычисления, конечно, дают p и q точно. Значения p и q из (13.3) бесполезны для определения λ_1 .

Вычисление комплексного собственного вектора

14. Если компоненты по $x_3, \dots, x_n$ уже исчезли, мы имеем

$$u_s = \alpha_1 x_1 + \bar{\alpha}_1 \bar{x}_1, \quad v_{s+1} = Au_s = \alpha_1 \lambda_1 x_1 + \bar{\alpha}_1 \bar{\lambda}_1 \bar{x}_1. \quad (14.1)$$

Если положить

$$\alpha_1 x_1 = z_1 + iw_1, \quad \lambda_1 = \xi_1 + i\eta_1, \quad (14.2)$$

то

$$u_s = 2z_1, \quad v_{s+1} = 2\xi_1 z_1 - 2\eta_1 w_1, \quad (14.3)$$

$$z_1 + iw_1 = \frac{1}{2} [u_s + i(\xi_1 u_s - v_{s+1})/\eta_1]. \quad (14.4)$$

С точностью до нормирующего множителя мы, следовательно, имеем

$$x_1 = \eta_1 u_s + i(\xi_1 u_s - v_{s+1}). \quad (14.5)$$

Если η_1 мало, u_s и v_{s+1} могут быть почти параллельны, и вектор x_1 плохо определен. Это можно хорошо проиллюстрировать также на примере § 13. Заметим, что (z_1, w_1) — инвариантное подпространство (гл. 3, § 63), так как из уравнения

$$A(z_1 + iw_1) = (\xi_1 + i\eta_1)(z_1 + iw_1) \quad (14.6)$$

следует

$$Az_1 = \xi_1 z_1 - \eta_1 w_1, \quad Aw_1 = \eta_1 z_1 + \xi_1 w_1. \quad (14.7)$$

Так как вещественным собственным значениям соответствуют вещественные собственные векторы, т. е. векторы с нулевой мнимой частью, можно было бы подумать, что если η_1 мало, то собственный вектор x_1 , нормированный так, что $\max(x_1) = 1 + i0$, будет иметь сравнительно небольшую мнимую часть. Однако это неверно, если только не существует матрицы, близкой к A и имеющей действительный квадратичный делитель, близкий к $(\lambda - \lambda_1)^2$. Например, нормированный собственный вектор x_1 , соответствующий матрице (13.1), равен

$$x_1^T = (-0,0417 - i0,9782, \quad 1 + i0). \quad (14.8)$$

Грубый результат, полученный для этой матрицы, не вызван нашим выбором $(1, 1)$ для u_s^T . Почти любой вектор дал бы такой же грубый результат.

Это становится ясным, если мы рассмотрим случай вещественной матрицы с вещественным доминирующим собственным значением, которому соответствуют два собственных вектора. Ясно, что мы не можем получить оба этих собственных вектора из одной последовательности $u_0, u_1, u_2, \dots$. Мы можем получить лишь один вектор из соответствующего подпространства. Это является предельной ситуацией, соответствующей комплексной паре собственных значений $(\xi_1 + i\eta_1)$, когда $\eta_1 \rightarrow 0$ и элементарные делители остаются линейными.

Сдвиг начала

15. Очевидно, что если мы итерируем с $(A - pI)$, используя какое-либо вещественное значение p , мы не можем «разделить» комплексно сопряженную пару собственных векторов. Если, однако, мы возьмем комплексное значение p и будем использовать итерации комплексных векторов, то модули $(\lambda_1 - p)$ и $(\bar{\lambda}_1 - p)$ будут различны и мы сможем получить сходимость к x_1 . В этом случае у нас будет приблизительно вдвое больше работы, но это оправдано, так как вместе с x_1 и λ_1 мы получаем и $\bar{x}_1$ и $\bar{\lambda}_1$. В этой связи следует заметить, что в методе § 12 мы пытались определить комплексно сопряженную пару собственных векторов, сделав лишь немного больше вычислений, чем для одного вещественного вектора.

Мой опыт с использованием комплексного сдвига не был особенно удачным, и я думаю, что лучше употреблять метод § 12, связанный лишь с вещественным сдвигом, но вычислять векторы u_s с двойной точностью. Так как элементы A определены с одинарной точностью, это потребует только вдвое больше арифметических действий, чем при использовании векторов с одинарной точностью, что совсем неплохо, так как мы получаем два собственных значения. Можно проверить, что если $|\eta_1| > 2^{-\frac{1}{2}t} \|A\|$, то векторы, вычисленные с двойной точностью, полностью предохраняют от плохой обусловленности, которую мы обсуждали, хотя для получения всех преимуществ мы должны продолжать итерации до тех пор, пока компоненты по оставшимся собственным векторам будут порядка 2^{-2t} вместо 2^{-t} .

Нелинейные делители

16. Поведение простого степенного метода в случае, когда A имеет нелинейные элементарные делители, иллюстрируется рассмотрением случая, когда A — простая жорданова подматрица $C_r(a)$ (гл. 1, § 7). Имеем

$$[C_r(a)]^s e_i = \binom{s}{i-1} a^{s-i+1} e_1 + \binom{s}{i-2} a^{s-i+2} e_2 + \dots \\ \dots + \binom{s}{1} a^{s-1} e_{i-1} + a^s e_i. \quad (16.1)$$

Ясно, что при $s \rightarrow \infty$ член с e_1 начинает доминировать, но сходимость медленная, так как правую часть (16.1) можно записать в виде

$$\binom{s}{i-1} a^{s-i+1} \left[e_1 + \frac{i-1}{s-i+2} a e_2 + \dots \right], \quad (16.2)$$

и выражение в скобках имеет асимптотику

$$e_1 + \frac{i-1}{s} a e_2. \quad (16.3)$$

Начиная с произвольного вектора u_0 , мы обязательно будем иметь сходимость к единственному собственному вектору e_1 , но асимптотическая сходимость будет медленнее, чем сходимость, соответствующая случаю различных собственных значений, как бы плохо ни были они разделены. Асимптотическая скорость сходимости может, однако, ввести в заблуждение; если a мало, мы видим, что член с e_2 в (16.3) мал. Это тривиальное замечание важно в связи с обратными итерациями (§ 53).

Возвращаясь к общему случаю, имеем

$$A^s = X C^s X^{-1}, \quad (16.4)$$

где C — жорданова каноническая матрица. Если доминирующее собственное значение A есть λ_1 и ему соответствует один делитель степени r , и если $C_r(\lambda_1)$ — первая жорданова подматрица в жордановой матрице C , то $C^s X^{-1} u_0 \sim k_s e_1$ для произвольных u_0 таких, что не все первые r компонент $X^{-1} u_0$ равны нулю. Следовательно,

$$A^s u_0 \sim k_s x_1. \quad (16.5)$$

Сходимость обычно слишком медленна для того, чтобы иметь практическое значение, но заметим, что компоненты в инвариантных подпространствах, соответствующих собственным значениям λ_i , меньшим по модулю, уменьшаются в $(\lambda_i/\lambda_1)^s$ раз после s итераций. Мы поэтому расстанемся с вектором, лежащим в подпространстве, соответствующем λ_1 , уже на сравнительно ранней стадии. Мы вернемся к нелинейным делителям в §§ 32, 41, 53.

Одновременное определение нескольких собственных значений

17. Основной чертой метода § 12 является то, что из одной последовательности итераций определяется два собственных значения. Тот факт, что собственные значения были комплексно сопряженными, на самом деле не играет существенной роли, и метод может быть обобщен для определения нескольких вещественных или комплексных собственных значений.

Предположим, например, что

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq |\lambda_4| \geq \dots \geq |\lambda_n|. \quad (17.1)$$

В итерированных векторах компоненты по $x_4, \dots, x_n$ будут быстро уменьшаться, и мы вскоре достигнем стадии, когда u_s будет в основном $\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$. Если мы определим p_2, p_1, p_0 уравнением

$$(\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) \equiv \lambda^3 + p_2 \lambda^2 + p_1 \lambda + p_0, \quad (17.2)$$

то

$$(A^3 + p_2 A^2 + p_1 A + p_0 I) u_s = 0, \quad (17.3)$$

что дает

$$- [A^3 u_s] = [A^2 u_s \mid A u_s \mid u_s] \begin{bmatrix} p_2 \\ p_1 \\ p_0 \end{bmatrix}. \quad (17.4)$$

Коэффициенты p_i можно таким образом получить из любых последовательных четырех итераций, следующих за u_s .

Использование подобной техники можно было бы рекомендовать, если бы $|\lambda_1|, |\lambda_2|, |\lambda_3|$ были близки, так как именно при таких обстоятельствах итерации медленно сходятся к доминирующему собственному вектору. Но если $\lambda_1, \lambda_2, \lambda_3$ одного знака и близки друг к другу, не только полином (17.2) плохо обусловлен, но и уравнения, определяющие p_i , также плохо обусловлены. По этой причине метод имеет малую практическую ценность. Метод Крылова (гл. 6, § 20) является его предельным случаем. Вообще для определения r собственных значений хорошо обусловленным способом нужно иметь r независимых систем итерированных векторов. Такие методы будут обсуждаться в §§ 36—39.

Комплексные матрицы

18. Если матрица A комплексная, то мы можем использовать простые итерации, но при этом потребуются в четыре раза больше работы при каждой итерации. Можно производить комплексные сдвиги, но отыскание величин таких сдвигов представляет значительные трудности. Вообще говоря, для получения доминирующего собственного вектора комплексной матрицы неудовлетворительно полагаться лишь на итерации с матрицами типа $(A - pI)$, и мы рекомендуем процедуру, описанную в канце § 62.

Исчерпывание

19. Мы уже видели, что если A — вещественная и имеет вещественные собственные значения, то мы можем применить вещественный сдвиг и получить сходимость либо к λ_1 , либо к λ_n . В случае комплексной матрицы мы можем в принципе, используя подходящий сдвиг, сделать доминирующими несколько собственных значений, но обычно это сделать очень трудно.

Естественно спросить, не можем ли мы использовать наше знание λ_1 и x_1 таким образом, чтобы получить другой собственный вектор без опасности сходимости снова к x_1 . Один класс таких методов состоит в переходе к матрице, обладающей теми же собственными значениями, кроме λ_1 . Будем называть такие методы методами *исчерпывания*.

Вероятно, простейший из них принадлежит Хотеллингу (1933). Он применим, если известны собственное значение λ_1 и соответствующий ему собственный вектор x_1 симметричной матрицы A_1 . Определим A_2 соотношением

$$x_1^T x_1 = 1, \quad A_2 = A_1 - \lambda_1 x_1 x_1^T. \quad (19.1)$$

Тогда из ортогональности собственных векторов имеем

$$A_2 x_i = A_1 x_i - \lambda_1 x_1 x_1^T x_i = \begin{cases} 0 & \text{при } i = 1, \\ \lambda_i x_i & \text{при } i \neq 1. \end{cases} \quad (19.2)$$

Следовательно, собственные значения A_2 равны нулю, $\lambda_2, \dots, \lambda_n$, и соответствующие собственные векторы равны $x_1, x_2, \dots, x_n$, так что доминирующее собственное значение λ_1 превратилось в нуль.

Если A_1 — несимметричная, то существует соответствующая техника исчерпывания, также принадлежащая Хотеллингу (1933), но она требует знания вместе с x_1 и левого собственного вектора y_1 . Если оба они определены и нормированы так, что $y_1^T x_1 = 1$, то, полагая

$$A_2 = A_1 - \lambda_1 x_1 y_1^T, \quad (19.3)$$

получим в силу биортогональности векторов x_i и y_i

$$A_2 x_i = A_1 x_i - \lambda_1 x_1 y_1^T x_i = \begin{cases} 0 & \text{при } i = 1, \\ \lambda_i x_i & \text{при } i \neq 1. \end{cases} \quad (19.4)$$

Оба эти метода исчерпывания имеют весьма плохую численную устойчивость, и я не рекомендую их использовать.

Исчерпывание, основанное на преобразованиях подобия

20. Рассмотрим исчерпывание, основанное на преобразованиях подобия A_1 . Пусть H_1 — любая неособенная матрица такая, что

$$H_1 x_1 = k_1 e_1, \quad (20.1)$$

где, очевидно, $k_1 \neq 0$. Имеем

$$A_1 x_1 = \lambda_1 x_1, \quad \text{откуда} \quad H_1 A_1 (H_1^{-1} H_1) x_1 = \lambda_1 H_1 x_1. \quad (20.2)$$

Уравнения (20.1) и (20.2) дают

$$H_1 A_1 H_1^{-1} e_1 = \lambda_1 e_1, \quad (20.3)$$

и, следовательно, первый столбец $H_1^* A_1 H_1^{-1}$ равен $\lambda_1 e_1$. Полагая

$$A_2 = H_1 A_1 H_1^{-1} = \begin{bmatrix} \lambda_1 & \vdots & b_1^T \\ 0 & \vdots & B_2 \end{bmatrix}, \quad (20.4)$$

получим, что матрица $(n-1)$ -го порядка B_2 имеет, очевидно, собственные значения $\lambda_2, \dots, \lambda_n$. Для того чтобы найти следующее собственное значение λ_2 , мы можем работать с B_2 . Если соответствующий собственный вектор B_2 обозначить через $x_2^{(2)}$, то, полагая

$$\begin{bmatrix} \lambda_1 & \vdots & b_1^T \\ 0 & \vdots & B_2 \end{bmatrix} \begin{bmatrix} \alpha \\ x_2^{(2)} \end{bmatrix} = \lambda_2 \begin{bmatrix} \alpha \\ x_2^{(2)} \end{bmatrix} \equiv \lambda_2 x_2^{(1)}, \quad (20.5)$$

получим

$$(\lambda_1 - \lambda_2)\alpha + b_1^T x_2^{(2)} = 0, \quad (20.6)$$

и окончательно из (20.4)

$$x_2 = H_1^{-1} x_2^{(1)}. \quad (20.7)$$

Таким образом, зная собственный вектор B_2 , мы можем определить соответствующий собственный вектор A_1 .

После того как мы нашли собственный вектор $x_2^{(2)}$ матрицы B_2 , мы подобным образом можем определить H_2 так, что

$$H_2 x_2^{(2)} = k_2 e_1. \quad (20.8)$$

Тогда

$$H_2 B_2 H_2^{-1} = \left[\begin{array}{c|c} \lambda_2 & b_2^T \\ \hline 0 & B_3 \end{array} \right], \quad (20.9)$$

где B_3 — матрица $(n-2)$ -го порядка, собственные значения которой равны $\lambda_3, \dots, \lambda_n$. По собственному вектору $x_3^{(3)}$ матрицы B_3 мы можем определить собственный вектор $x_3^{(2)}$ матрицы B_2 , затем вектор $x_3^{(1)}$ матрицы $H_1 A_1 H_1^{-1}$ и, наконец, вектор x_3 матрицы A_1 . Продолжая этот процесс, можем получить все собственные значения и собственные векторы A_1 .

Если мы обозначим

$$K_r = \left[\begin{array}{c|c} I_{r-1} & 0 \\ \hline 0 & H_r \end{array} \right], \quad (20.10)$$

то

$$K_r \dots K_3 K_2 K_1 A_1 K_1^{-1} K_2^{-1} K_3^{-1} \dots K_r^{-1} = \left[\begin{array}{c|c} T_r & C_r \\ \hline 0 & B_{r+1} \end{array} \right], \quad (20.11)$$

где T_r суть верхняя треугольная матрица с диагональными элементами $\lambda_1, \dots, \lambda_r$, а B_{r+1} имеет собственные значения, равные $\lambda_{r+1}, \dots, \lambda_n$. Последовательное исчерпывание дает, следовательно, приведение к треугольному виду с помощью преобразований подобия (гл. 1, §§ 42, 47).

Исчерпывание, использующее инвариантные подпространства

21. Прежде чем рассматривать детально процесс исчерпывания, обсудим случай, когда мы определили инвариантное подпространство матрицы A (гл. 3, § 63)*), но не обязательно отдельные собственные векторы. Предположим, что мы имеем матрицу X , с m линейно независимыми столбцами, порождающими инвариантное подпространство. Тогда

$$AX = XM, \quad (21.1)$$

где M — некоторая $(m \times m)$ -матрица.

Предположим также, что H — неособенная матрица такая, что

$$HX = \left[\begin{array}{c} T \\ \hline 0 \end{array} \right], \quad (21.2)$$

*) В дальнейшем иногда мы будем говорить вместо «подпространство, натянутое на векторы $v_1, v_2, \dots, v_r$ », короче — «подпространство $(v_1 \mid v_2 \mid \dots \mid v_r)$ » и даже «подпространство v », где v — матрица, составленная из столбцов $v_1, v_2, \dots, v_r$. (Прим. перев.)

где T — неособенная $(m \times m)$ -матрица (обычно на практике верхняя треугольная). Если мы положим

$$HAN^{-1} = \begin{bmatrix} B & C \\ D & E \end{bmatrix}, \quad (21.3)$$

то из соотношения

$$HAN^{-1}HX = HXM \quad (21.4)$$

получим

$$\begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} T \\ 0 \end{bmatrix} = \begin{bmatrix} T \\ 0 \end{bmatrix} M, \quad (21.5)$$

что дает

$$BT = TM \quad \text{и} \quad DT = 0. \quad (21.6)$$

По второму из этих соотношений $D = 0$, и следовательно, собственные значения A совпадают с собственными значениями B и E . Из первого соотношения (21.6) следует, что собственные значения B совпадают с собственными значениями M .

Исчерпывание при помощи устойчивых элементарных преобразований

22. Вернемся теперь к случаю, когда мы знаем один собственный вектор x_1 и хотим определить матрицу H_1 , удовлетворяющую (20.1). Мы можем определить H_1 устойчивым способом как произведение вида $N_1 I_{1,1'}$, причем считаем, что максимальный по модулю элемент x_1 находится в позиции $1'$. В предположении, что x_1 нормирован так, что $\max(x_1) = 1$, процесс исчерпывания принимает следующий вид:

- (i) Переставим строки 1 и $1'$ и столбцы 1 и $1'$ матрицы A_1 .
- (ii) Переставим элементы 1 и $1'$ вектора x_1 и назовем полученный вектор y_1 .
- (iii) Для всех значений i от 2 до n вычислим:

$$i\text{-я строка} = i\text{-я строка} - (y_i \times 1\text{-я строка}).$$

- (iv) Для всех значений i от 2 до n вычислим:

$$1\text{-й столбец} = 1\text{-й столбец} + (y_i \times i\text{-й столбец}).$$

На самом деле шаг (iv) не нужно выполнять, так как мы знаем, что 1-й столбец должен стать $(\lambda_1, 0, \dots, 0)$. Таким образом, процесс исчерпывания требует только $(n-1)^2$ умножений. Заметим, что при таком исчерпывании b_1^T из уравнения (20.4) состоит из элементов $1'$ -й строки матрицы A_1 , за исключением элемента $a_{1',1'}$, переставленного очевидным образом. Заметим также, что даже если A_1 симметрична, это, вообще говоря, неверно для A_2 .

23. Пусть мы имеем «инвариантное подпространство» X , как в § 21, и хотим определить H , удовлетворяющую (21.2). Опять это можно осуществить весьма удобно, используя устойчивые элементарные преобразования. Соответствующая H может быть определена как произведение

$$N_m I_{m,m'} \dots N_2 I_{2,2'} N_1 I_{1,1'}, \quad (23.1)$$

в котором матрицы совпадают с теми, которые можно было бы получить, осуществляя гауссово исключение с перестановками в матрице X . (Тот факт, что она имеет лишь m столбцов вместо n , несуществен.) Очевидно,

мы имеем

$$N_m I_{m,m'} \dots N_2 I_{2,2'} N_1 I_{1,1'} X = \begin{bmatrix} T \\ 0 \end{bmatrix}, \quad (23.2)$$

где T — верхняя треугольная. Если перестановки не необходимы, H имеет вид $\begin{bmatrix} L & 0 \\ M & I_{n-m} \end{bmatrix}$, и если X^T есть $[U : V]$, то (23.2) эквивалентно равенствам

$$LU = T \quad \text{и} \quad MU + V = 0. \quad (23.3)$$

Вычисление HAH^{-1} производится в m шагов, причем в r -ом из них осуществляется преобразование подобия с матрицей $N_r I_{r,r'}$. Заметим, что нам не надо вычислять часть матрицы, обозначенную D в (21.5), так как известно, что она должна быть нулем. Так как мы имеем полную информацию о перестановках перед началом процесса исчерпывания, то знаем, из каких строк и столбцов получается подматрица D . Легко организовать вычисления так, чтобы скалярные произведения вычислялись накоплением, как при определении H (аналогично треугольному разложению с перестановками), так и при вычислении исчерпанной матрицы.

24. Наиболее часто ситуация использования инвариантных подпространств возникает в связи с комплексной парой собственных значений. Если мы используем метод § 12, то после того, как сходимость уже имеет место, любые два последовательных u_s и u_{s+1} определяют инвариантное подпространство.

Если определен комплексный собственный вектор $x_1 = z_1 + iw_1$, процесс исчерпывания имеет ряд интересных черт. Предположим, что $z_1 + iw_1$ нормирован так, что $\max(x_1) = 1 + i0$. Пусть, например, имеем при $n = 5$ и $1' = 3$

$$\left. \begin{aligned} z_1^T &= (a_1, a_2, 1, a_4, a_5), \\ w_1^T &= (b_1, b_2, 0, b_4, b_5). \end{aligned} \right\} \quad (24.1)$$

Первая перестановка дает

$$\begin{bmatrix} 1 & a_2 & a_1 & a_4 & a_5 \\ 0 & b_2 & b_1 & b_4 & b_5 \end{bmatrix} = \begin{bmatrix} 1 & a'_2 & a'_3 & a'_4 & a'_5 \\ 0 & b'_2 & b'_3 & b'_4 & b'_5 \end{bmatrix}, \quad (24.2)$$

и в матрице N_1 имеем $n_{j1} = a'_j$. Приведение первого столбца x оставляет второй неизменным; он потом независимо приводится при помощи умножения слева на $N_2 I_{2,2'}$. Итог таков, что мы трактуем z_1 как действительный собственный вектор и осуществляем одно исчерпывание, а затем считаем w_1 (с опущенным нулевым элементом) действительным собственным вектором исчерпанной матрицы и соответственно осуществляем следующее исчерпывание.

Исчерпывание при помощи унитарных преобразований

25. Для выполнения исчерпывания можно использовать и унитарные преобразования. В случае одного собственного вектора в качестве матрицы H_1 в уравнении (20.4) можно взять либо одну матрицу отражения $(I - 2ww^T)$, где, вообще говоря, w не имеет нулевых компонент, либо произведение вращений в плоскостях $(1, 2)$, $(1, 3)$, ..., $(1, n)$. Если x_1 — вещественный, то эти матрицы ортогональны.

Для исчерпывания нужно приблизительно $8n^2$ умножений, если мы применяем плоские вращения, и $4n^2$, если мы используем матрицы отра-

жения. Это дает обычное отношение два к одному. Однако заметим, что с матрицами отражения мы имеем в четыре раза больше умножений, чем при неунитарных преобразованиях; в предыдущих случаях это отношение было два к одному. Это случилось потому, что при исчерпывании при помощи неунитарных преобразований нам на самом деле не надо выполнять умножение справа на H_1^{-1} . Это умножение всегда оставляет последние $(n - 1)$ столбцов неизменными, и мы можем положить первый столбец равным $\lambda_1 e_1$ без вычислений. Хотя в случае унитарных преобразований мы также можем взять окончательный первый столбец равным $\lambda_1 e_1$, это не спасет нас от вычислений, так как все остальные столбцы меняются при умножении справа.

Если исходная матрица была эрмитова, то после исчерпывания матрица будет также эрмитовой и, следовательно, будет иметь вид

$$\begin{bmatrix} \lambda_1 & 0 \\ 0 & B_2 \end{bmatrix}, \quad (25.1)$$

где B_2 — эрмитова. Исчерпывание может быть проведено почти так же, как первый шаг метода Хаусхолдера (правда, w уже не имеет нулевого первого элемента), включая способ, описанный в гл. 5, § 30, для упрощения в симметричном случае. Интересно заметить, что этот метод исчерпывания был впервые описан Феллером и Форсайтом (1951), хотя и в сильно отличных терминах. К сожалению, он не привлек к себе внимания (может быть, потому что он был лишь одним из большого числа методов исчерпывания, описанных в статье), хотя он мог бы привести к значительно более раннему широкому использованию матриц отражения в численном анализе.

Аналогично мы можем работать с инвариантным подпространством, натянутым на m столбцов, преобразуя их так же, как мы преобразовывали первые m столбцов квадратной матрицы при ортогональном приведении к треугольной (гл. 4, § 46). Это требует m матриц отражения с векторами w_i , имеющими соответственно 0, 1, 2, ..., $m - 1$ нулевых элементов. Соответствующая матрица T снова верхняя треугольная. При комплексном собственном векторе $z_1 + iw_1$ вещественной матрицы мы используем вещественное инвариантное подпространство $(z_1 | w_1)$, так что достаточно вещественных (ортогональных) матриц, и комплексная арифметика не привлекается.

Численная устойчивость

26. Методы исчерпывания §§ 22 и 25 исключительно стабильны. Предположим, что u_s принят в качестве собственного вектора и $\lambda = \max (Au_s)$ — в качестве собственного значения. Мы можем опустить индекс s и для простоты предположить, что наибольшим по модулю элементом u_s является первый. Положим

$$Au - \lambda u = \eta \quad (26.1)$$

и предположим, что $\|\eta\|$ мала по сравнению с $\|A\|$. Однако (53.6) гл. 3 и § 10 гл. 2 показывают, что это не означает, что u обязательно точен. Действительно, если существует более чем одно собственное значение, близкое к λ , то u может быть очень неточным собственным вектором. Если мы положим

$$A_1 = \begin{bmatrix} a_{11} & a^T \\ c & B_1 \end{bmatrix}, \quad u^T = (1 | v^T), \quad \text{где} \quad |v_i| \leq 1, \quad (26.2)$$

то

$$H_1 = \left[\begin{array}{c|c} 1 & 0 \\ \hline -v & I \end{array} \right]. \quad (26.3)$$

Используя обозначение A_2 для *вычисленной* матрицы после исчерпывания, имеем

$$A_2 = \left[\begin{array}{c|c} \lambda & a^T \\ \hline 0 & B_2 \end{array} \right], \quad (26.4)$$

где на самом деле вычислены лишь элементы B_2 . При вычислениях с фиксированной запятой, например, имеем

$$a_{ij}^{(2)} = a_{ij}^{(1)} - u_i a_{1j}^{(1)} + \varepsilon_{ij}, \quad |\varepsilon_{ij}| \leq \frac{1}{2} 2^{-t} \quad (1 < i, j \leq n), \quad (26.5)$$

что дает

$$B_2 = B_1 - va^T + F, \quad (26.6)$$

где F — матрица с элементами ε_{ij} . Рассмотрим теперь матрицу $\tilde{A}_1$, определенную по вычисленной матрице A_2 следующим образом:

$$\begin{aligned} \tilde{A}_1 &= H_1^{-1} A_2 H_1 = \left[\begin{array}{c|c} \lambda - a^T v & a^T \\ \hline \lambda v - B_2 v - va^T v & B_2 + va^T \end{array} \right] = \\ &= \left[\begin{array}{c|c} \lambda - a^T v & a^T \\ \hline (\lambda I - B_1 - F) v & B_1 + F \end{array} \right]. \end{aligned} \quad (26.7)$$

Вычисленная матрица A_2 имеет в точности те же самые собственные значения, что и $\tilde{A}_1$, и очевидно, что A_1 имеет λ и u в качестве *точных* собственного значения и собственного вектора. Мы можем поэтому сказать, что вычисленная матрица A_2 получена в результате нахождения *точного* собственного значения и *точного* собственного вектора матрицы $\tilde{A}_1$ с последующим выполнением *точного* процесса исчерпывания. Мы хотим сравнить $\tilde{A}_1$ и A_1 . Обозначим $\eta^T = (\eta_1 \mid \zeta)$. Тогда (26.1) дает

$$a_{11} + a^T v - \lambda = \eta_1, \quad c + B_1 v - \lambda v = \zeta, \quad (26.8)$$

и, следовательно,

$$\tilde{A}_1 = \left[\begin{array}{c|c} a_{11} - \eta_1 & a^T \\ \hline c - \zeta - Fv & B_1 + F \end{array} \right], \quad (26.9)$$

$$\tilde{A}_1 - A_1 = \left[\begin{array}{c|c} -\eta_1 & 0 \\ \hline -\zeta - Fv & F \end{array} \right], \quad (26.10)$$

что дает, например,

$$\|\tilde{A}_1 - A_1\|_\infty \leq \|\eta\|_\infty + \frac{1}{2} (n-1) 2^{-t}. \quad (26.11)$$

Предполагая, что невязка η мала, мы можем быть уверены, что собственные значения A_2 будут близки к собственным значениям A_1 (если они хорошо обусловлены), хотя A_2 может сильно отличаться от матрицы, которая получилась бы, если бы u был точным.

Если мы рассмотрим последовательность исчерпываний, то существует опасность, что элементы последовательных A_i будут постоянно возрастать

по величине и A_i могут становиться все более плохо обусловленными. Мы уже обращали внимание на такую возможность раньше, когда рассматривалась последовательность устойчивых неунитарных преобразований. В этом частном случае опасность, как кажется, меньше, чем обычно. Действительно, на последовательных этапах исчерпывания мы работаем только с матрицами B_r из нижнего правого угла A_r , и обычно обусловленность B_r улучшается с возрастанием r , даже если это неверно для A_r . В качестве предельного примера рассмотрим A_r , имеющую квадратичный делитель. Если пренебречь ошибками округления, все A_r должны иметь квадратичный делитель, но если мы по одному собственному значению найдем соответствующую B_r , то она уже не имеет квадратичного делителя. Следовательно, обусловленность оставшегося из пары собственных значений должна улучшиться!

Численный пример

27. В табл. 1 мы показываем исчерпывание матрицы третьего порядка. Из приведенных здесь собственных значений и собственных векторов видно, что λ_1 и λ_2 близки. Вектор u_s принят в качестве собственного вектора и дает точное собственное значение и малый вектор невязки, несмотря на то, что он имеет ошибки во втором знаке. Вычисленная исчерпанная матрица имеет значительные расхождения с матрицей, которая получилась бы при точных вычислениях, но тем не менее оба λ_2 и λ_3 верны с рабочей точностью. Вектор $\tilde{u}$ является точным собственным вектором $\tilde{A}_1$, и точное исчерпывание $\tilde{A}_1$ дает вычисленную A_2 . Следовательно, собственные значения вычисленной A_2 совпадают с собственными значениями $\tilde{A}_1$; заметим, что $\tilde{A}_1$ и A_1 очень близки.

Таблица 1

A_1				x_1	x_2	x_3
$\begin{bmatrix} 0,987 & 0,400 & -0,487 \\ -0,079 & 0,500 & 0,479 \\ 0,082 & 0,400 & 0,418 \end{bmatrix}$			$\lambda_1=0,905$ $\lambda_2=0,900$ $\lambda_3=0,100$	1,000 0,143 0,286	1,0 1,0 1,0	1,0 -1,0 1,0

$$u_s^T = [1,000 \quad 0,200 \quad 0,333] = \text{вычисленный } x_1.$$

$$\eta_s^T = 10^{-3}[-0,171 \quad -0,498 \quad -0,171] = \text{вектор невязки.}$$

$$\text{Вычисленная } A_2 = \begin{bmatrix} 0,905 & 0,400 & -0,487 \\ 0 & 0,420 & 0,576 \\ 0 & 0,267 & 0,580 \end{bmatrix} \quad \begin{matrix} \lambda'_2=0,900 \\ \lambda'_1=0,100 \end{matrix}$$

$$\text{Точная } A_2 = \begin{bmatrix} 0,905 & 0,400 & -0,487 \\ 0 & 0,443 & 0,549 \\ 0 & 0,286 & 0,557 \end{bmatrix}$$

$$\tilde{A}_1 = \begin{bmatrix} 0,987\,171 & 0,400\,000 & -0,487\,000 \\ -0,078\,373\,800 & 0,500\,000 & 0,478\,600 \\ 0,082\,187\,943 & 0,400\,200 & 0,417\,829 \end{bmatrix}$$

$$\text{Собственный вектор вычисленной } B_2 = [1,000; 0,833] = [x_2^{(2)}]^T$$

$$\text{Вычисленная } A_3 = \begin{bmatrix} 0,905 & 0,400 & -0,487 \\ 0 & 0,900 & 0,576 \\ 0 & 0 & 0,100 \end{bmatrix}$$

Вычисленные	x_1	x_2	x_3
	1,000	0,750	1,000
	0,200	1,000	-1,000
	0,333	0,958	1,000

Второе собственное значение определяется из подматрицы второго порядка B_2 вычисленной A_2 . Так как она имеет два хорошо отделенных собственных значения, проблема нахождения ее собственных векторов хорошо обусловлена, и, следовательно, итерации ведут к точному собственному вектору B_2 . Далее дана вычисленная A_3 и три вычисленных собственных вектора матрицы A . Вычисленный x_2 неточен, но это и ожидалось; его неточность нельзя полностью относить к процессу исчерпывания. Он не менее точен, чем вычисленный x_1 , и это было получено использованием исходной матрицы. Однако вычисленные x_1 и x_2 порождают свое подпространство точно. В самом деле, вычисленный $x_2^{(2)}$ согласуется с точным $x_2^{(2)}$ с рабочей точностью. Это следует из того, что вектор из подпространства x_1, x_2 , имеющий нулевую первую компоненту, хорошо определен. С другой стороны, x_3 , полученный, используя «очень неточное» исчерпывание, верен с принятой точностью. Это можно было ожидать, так как x_3 — хорошо обусловленный собственный вектор, и ошибки, сделанные в первом исчерпывании, эквивалентны работе с $\tilde{A}_1$ вместо A_1 .

Для получения x_2 по собственному вектору B_2 имеем

$$x_2 = \begin{bmatrix} 1,000 \\ 0,200 \\ 0,333 \end{bmatrix} + \alpha \begin{bmatrix} 0 \\ 1,000 \\ 0,833 \end{bmatrix}, \quad (27.1)$$

где α определяется приравниванием первых элементов Ax_2 и $\lambda_2 x_2$. Это дает

$$\lambda_2 = \lambda_1 + \alpha(-0,005671), \quad \text{или} \quad -0,005 = \alpha(-0,005671).$$

Заметим, что α определяется как отношение двух малых величин. Это неизбежно, если A_1 имеет близкие собственные значения, но не близка к матрице, имеющей нелинейные делители. В предельном случае, когда λ_1 и λ_2 равны, но A имеет линейные делители, коэффициент при α (т. е. $a^T x_2^{(2)}$) в точности равен нулю, что показывает, что α произвольно, как мы и могли ожидать. Если A_1 имеет квадратичный делитель, то $a^T x_2^{(2)}$ не нуль, и, следовательно, $x_2 = x_1$ и мы получаем только один собственный вектор. Вектор $x_2^{(2)}$ дает линейно независимый вектор в инвариантном подпространстве A_1 , соответствующем квадратичному делителю.

Устойчивость унитарных преобразований

28. Как можно ожидать, унитарные исчерпывания § 25 также очень устойчивы и даже не страдают от опасности роста по величине элементов последовательно исчерпываемых матриц и ухудшения их обусловленности. Это не новый факт в анализе. Мы снова заметим, что λ и u суть точные собственное значение и собственный вектор $(A_1 - \eta u^T)$, и общий анализ главы 3 непосредственно приложим. Этот анализ в случае вычислений с фиксированной запятой и накоплением скалярных произведений был в деталях проведен Уилкинсоном (1962b).

Для исчерпывания, использующего матрицы отражения с накоплением с плавающей запятой, анализ, приведенный в гл. 3, § 45, дает, что вычисленная A_r в точности подобна $(A_1 + F_r)$, где

$$\|F_r\|_E \leq \| \eta_1 \|_2 + \| \eta_2 \|_2 + \dots + \| \eta_{r-1} \|_2 + 2(r-1)(1+x)^{2r-4} x \| A_1 \|_E. \quad (28.1)$$

Здесь $x = (12,36) 2^{-t}$, а η_i — вектор невязки на i -м шаге. Следовательно, если мы итерируем до тех пор, пока вектор невязки не станет мал,

собственные значения сохраняются (в той мере, насколько они хорошо обусловлены), даже если принятые собственные векторы плохи.

Очевидно, что симметрия A_1 сохраняется. Некоторые трудности в анализе возникают из-за того, что $(A_1 - \eta u^T)$, где u^T — точный собственный вектор, не будет, вообще говоря, симметричной. Однако если мы возьмем λ равным отношению Релея (гл. 3, § 54) для u , то $\eta^T u = 0$ и, следовательно,

$$(A_1 - \eta u^T - u \eta^T) u = \lambda u. \quad (28.2)$$

Матрица в скобках уже будет в точности симметричной.

Мы знаем, пренебрегая ошибками округления, что индивидуальные числа обусловленности собственных значений A_i инвариантны при унитарном исчерпывании, но это не совсем то, что нужно. На каждом этапе мы работаем не с целой матрицей A_i , а только с ее подматрицей B_i , где

$$A_i = \left[\begin{array}{c|c} T_{i-1} & C_{i-1} \\ \hline 0 & B_i \end{array} \right], \quad (28.3)$$

а T_{i-1} — верхняя треугольная матрица, имеющая вычисленные собственные значения в качестве диагональных элементов. Если λ_k — собственное значение B_i , легко показать, что индивидуальное число обусловленности $\tilde{s}_k$ собственного значения λ_k по отношению к B_i не меньше по модулю s_k — числа обусловленности λ_k по отношению к A_i , и, следовательно, к A_1 . Действительно, если

$$\left[\begin{array}{c|c} T_{i-1} & C_{i-1} \\ \hline 0 & B_i \end{array} \right] \begin{bmatrix} u \\ v \end{bmatrix} = \lambda_k \begin{bmatrix} u \\ v \end{bmatrix}, \quad \left[\begin{array}{c|c} T_{i-1}^T & 0 \\ \hline C_{i-1}^T & B_i^T \end{array} \right] \begin{bmatrix} 0 \\ w \end{bmatrix} = \lambda_k \begin{bmatrix} 0 \\ w \end{bmatrix}, \quad (28.4)$$

то

$$B_i v = \lambda_k v, \quad B_i^T w = \lambda_k w, \quad (28.5)$$

что показывает, что $|\tilde{s}_k| \geq |s_k|$. Следовательно, вообще говоря, собственные значения B_i менее чувствительны, чем собственные значения A_i . Повышенная чувствительность собственных значений A_i является следствием возможных возмущений остальных трех подматриц. Мы не имеем таких хороших результатов в случае неунитарного исчерпывания, хотя эксперименты на АСЕ показали, что числа обусловленности *обычно* показывают постоянное улучшение.

Число умножений при неунитарном исчерпывании равно приблизительно n^2 по сравнению с $4n^2$ при унитарном исчерпывании, в то время как одна итерация требует приблизительно n^2 умножений, если A_1 — полная матрица. Если требуется значительное количество итераций для нахождения каждого собственного вектора, то работа, связанная с исчерпыванием, составляет малую часть общей работы. В этом случае аргументы в пользу использования унитарных преобразований с их безусловной гарантией устойчивости очень сильны.

Исчерпывание при помощи неподобных преобразований

29. Для случая, когда мы определили один собственный вектор x_1 , Виландт (1944b) описал общую форму исчерпывания, которое не является преобразованием подобия. Предположим, что u_1 — какой-либо вектор такой, что

$$u_1^T x_1 = \lambda_1. \quad (29.1)$$

Рассмотрим матрицу

$$A_2 = A - x_1 u_1^T. \quad (29.2)$$

Имеем

$$A_2 x_1 = A x_1 - x_1 u_1^T x_1 = \lambda_1 x_1 - \lambda_1 x_1 = 0, \quad (29.3)$$

и, следовательно, x_1 все еще является собственным вектором, но соответствующее собственное значение равно нулю. С другой стороны,

$$\begin{aligned} A_2(x_i - \alpha_i x_1) &= A x_i - \alpha_i A x_1 - x_1 u_1^T x_i + \alpha_i x_1 u_1^T x_1 = \\ &= \lambda_i x_i - x_1 u_1^T x_i = \lambda_i [x_i - (u_1^T x_i / \lambda_i) x_1] \quad (i = 2, \dots, n). \end{aligned} \quad (29.4)$$

Следовательно, выбирая $\alpha_i = (u_1^T x_i / \lambda_i)$, видим, что λ_i остается собственным значением, и соответствующий собственный вектор равен

$$x_i^{(2)} = x_i - (u_1^T x_i / \lambda_i) x_1. \quad (29.5)$$

Заметим, что n векторов $x_1, x_i^{(2)}$ ($i \geq 2$) линейно независимы. Действительно, если

$$\beta_1 x_1 + \sum_{i=2}^n \beta_i [x_i - (u_1^T x_i / \lambda_i) x_1] = 0, \quad (29.6)$$

то $\beta_i = 0$ ($i = 2, \dots, n$), так как x_i линейно независимы, и, следовательно, β_1 также равен нулю.

Исчерпывание Хотеллинга (§ 19) является частным случаем этого, в котором u_1 взято равным $\lambda_1 y_1$, где y_1 — левый собственный вектор, нормированный так, что $y_1^T x_1 = 1$. Так как $y_1^T x_i = 0$ ($i \neq 1$), собственные векторы в этом случае остаются прежними.

30. В общем процессе мы мысленно предполагали, что $\lambda_i \neq 0$, и хотя этого можно всегда достигнуть при помощи сдвига, это неприятная черта метода. Рассмотрим следующую матрицу:

$$A = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}, \quad \lambda_1 = 4, \quad \lambda_2 = 0, \quad x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}. \quad (30.1)$$

Вектор $u^T = (4, 0)$ подходит, и мы имеем

$$A_2 = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} [4, 0] = \begin{bmatrix} -2 & 1 \\ -4 & 2 \end{bmatrix}. \quad (30.2)$$

Матрица A_2 имеет квадратичный делитель λ^2 и гораздо более плохо обусловлена, чем A .

Предположим теперь, что v_1 — некоторый вектор, для которого

$$v_1^T x_1 = 1. \quad (30.3)$$

Имеем $v_1^T A x_1 = v_1^T \lambda_1 x_1 = \lambda_1$, так что $v_1^T A$ — подходящий вектор в качестве u_1^T в (29.1). Если мы возьмем этот вектор, то процесс Виландта дает

$$A_2 = A - x_1 v_1^T A. \quad (30.4)$$

При этом собственные векторы таковы:

$$x_i - (v_1^T A x_i / \lambda_i) x_1 = x_i - (v_1^T x_i) x_1 \quad (i = 2, \dots, n), \quad (30.5)$$

и нулевое значение для λ_i не играет теперь особой роли. Можно прямо проверить, что $x_i - (v_1^T x_i) x_1$ ($i = 2, \dots, n$) суть действительно собственные векторы A_2 . Ясно, что это не преобразование подобия, так как, вообще говоря, λ_1 меняется.

Если x_1 нормирован так, что $\max(x_i) = 1$, и если единичный элемент находится в позиции $1'$, естественным выбором для v_1 является вектор $e_{1'}$. Тогда

$$A_2 = A - x_1 e_1^T A. \quad (30.6)$$

Ясно, что строка $1'$ матрицы A_2 нулевая и i -я строка матрицы A_2 получена из i -й строки матрицы A вычитанием строки $1'$, умноженной на x_i . Следовательно, $I_{1,1'} A_2 I_{1,1'}$ имеет те же самые элементы в матрице $(n-1)$ -го порядка из правого нижнего угла, что и матрица при исчерпывании в § 22. Так как A_2 имеет нулевую строку $1'$, все собственные векторы, соответствующие λ_i ($i = 2, \dots, n$), имеют нулевую компоненту в позиции $1'$. С другой стороны, это очевидно и из того, что эти собственные векторы A_2 имеют вид $x_i - (e_1^T, x_i) x_1$. Ясно, что мы можем опустить $1'$ -й столбец и $1'$ -ю строку при итерировании с A_2 и работать только с матрицей $(n-1)$ -го порядка.

Дальнейшее обсуждение упростится, если мы предположим, что $1' = 1$, ибо необходимые изменения, если $1' \neq 1$, тривиальны. Исчерпанная матрица A_2 тогда имеет вид

$$A_2 = A - x_1 e_1^T A = (I - x_1 e_1^T) A. \quad (30.7)$$

Обозначим собственные векторы A_2 через $x_1, x_2^{(2)}, x_3^{(2)}, \dots, x_n^{(2)}$, где $x_i^{(2)}$ ($i \geq 2$) имеют нулевую первую компоненту. Если $x_2^{(2)}$ уже найден, мы можем выполнить второе исчерпывание и снова для простоты будем считать, что $2' = 2$. Следовательно,

$$A_3 = A_2 - x_2^{(2)} e_2^T A_2 = (I - x_2^{(2)} e_2^T) (I - x_1 e_1^T) A \quad (30.8)$$

и A_3 имеет нулевые строки 1 и 2.

Вообще говоря, процесс исчерпывания удобно осуществлять явно для получения матриц $A_2, A_3, \dots$ или по крайней мере соответствующих подматриц порядка $n-1, n-2, \dots$, которые используются при итерациях. Однако если матрица A имеет много нулей (редкая), то это делать не экономно, так как последующие матрицы уже не будут, вообще говоря, иметь много нулей. Если требуется определить лишь несколько собственных векторов такой матрицы, то обычно лучше работать с исчерпанными матрицами в факторизованной форме типа (30.8). Например, мы знаем, что при итерировании с A_3 любой собственный вектор имеет равные нулю две первые компоненты. Следовательно, мы можем взять начальный вектор u_0 в такой форме, и это будет верно для всех u_s . Полагая $Au_s = p_{s+1}$, вычислим

$$q_{s+1} = (I - x_1 e_1^T) p_{s+1} = p_{s+1} - (e_1^T p_{s+1}) x_1. \quad (30.9)$$

Ясно, что q_{s+1} получается из p_{s+1} вычитанием x_1 , умноженного на соответствующий множитель такой, чтобы первая компонента исчезла.

Имеем

$$(I - x_2^{(2)} e_2^T) q_{s+1} = q_{s+1} - (e_2^T q_{s+1}) x_2^{(2)} \equiv v_{s+1}, \quad (30.10)$$

так что v_{s+1} — это ненормированный вектор, полученный из q_{s+1} вычитанием $x_2^{(2)}$, умноженного на такой множитель, чтобы исчезла вторая компонента. Этот процесс можно использовать и в общем случае, когда мы

имеем A_r . Например, если $n = 6$ и $r = 4$, то

$$[x_1 | x_2^{(2)} | x_3^{(3)}] = \begin{bmatrix} 1 & 0 & 0 \\ \times & 1 & 0 \\ \times & \times & 1 \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}, \quad u_s = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \times \\ \times \end{bmatrix}, \quad Au_s = \begin{bmatrix} \times \\ \times \\ \times \\ \times \\ \times \\ \times \end{bmatrix}, \quad v_{s+1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \times \\ \times \\ \times \end{bmatrix}, \quad (30.11)$$

и Au_s нужно привести к виду v_{s+1} при помощи процесса исключения, который работает со столбцами, а не со строками. Вектор v_{s+1} затем нормируется и дает u_{s+1} . Мы увидим в § 34, что это почти точно процесс ступенчатых итераций Ф. Л. Бауэра.

Общее приведение, использующее инвариантные подпространства

31. Если X — инвариантное подпространство с m столбцами, так что

$$AX = XM, \quad (31.1)$$

где M — матрица m -го порядка, и если P — любая $(n \times m)$ -матрица такая, что

$$P^T X = M, \quad (31.2)$$

то обобщенная форма исчерпывания Виландта получится при

$$A_2 = A - XP^T. \quad (31.3)$$

Очевидно,

$$A_2 X = AX - XP^T X = XM - XM = 0, \quad (31.4)$$

так что все m линейно независимых столбцов X являются собственными векторами A_2 , соответствующими собственному значению, равному нулю. (Это верно, даже если X соответствует элементарному делителю степени m). Остальные собственные значения A_2 те же, что и у A . Покажем это. Доказательство следует из простого равенства

$$\det(I_n - RS) = \det(I_m - SR), \quad (31.5)$$

которое справедливо при произвольных матрицах R и S вида $(n \times m)$ и $(m \times n)$ соответственно. Равенство (31.5) является очевидным следствием того факта, что ненулевые собственные значения RS и SR совпадают (гл. 1, § 51). Далее, из (31.1) следует, что

$$\lambda X - AX = \lambda X - XM, \quad (31.6)$$

и потому

$$(\lambda I_n - A)X = X(\lambda I_m - M) \quad \text{или} \quad X(\lambda I_m - M)^{-1} = (\lambda I_n - A)^{-1}X. \quad (31.7)$$

Далее, используя (31.2), (31.5) и (31.7), получаем

$$\begin{aligned} \det(\lambda I_n - A_2) &= \det(\lambda I_n - A + XP^T) = \\ &= \det(\lambda I_n - A) \det[I_n + (\lambda I_n - A)^{-1}XP^T] = \\ &= \det(\lambda I_n - A) \det[I_n + X(\lambda I_m - M)^{-1}P^T] = \\ &= \det(\lambda I_n - A) \det[I_m + P^T X(\lambda I_m - M)^{-1}] = \\ &= \det(\lambda I_n - A) \det[I_m + M(\lambda I_m - M)^{-1}] = \\ &= \det(\lambda I_n - A) \lambda^m \det(\lambda I_m - M)^{-1}, \end{aligned} \quad (31.8)$$

что показывает, что собственные значения A_2 совпадают с собственными значениями A , за исключением принадлежащих M , которые заменены нулями.

32. Хаусхолдер (1961) нашел соотношения между инвариантными подпространствами A_2 и A . Эти соотношения являются обобщениями равенства (29.5) и страдают недостатком, аналогичным присутствию λ_i в знаменателе. Мы рассмотрим лишь случай, используемый на практике, для которого можно получить несколько более строгие результаты и более просто, чем в общем случае.

Пусть Q — некоторая матрица, для которой

$$Q^T X = I_m. \quad (32.1)$$

(Это естественное обобщение (30.3)). Тогда из (31.1) имеем

$$Q^T A X = Q^T X M = M, \quad (32.2)$$

и следовательно, $Q^T A$ это подходящая матрица P^T в равенстве (31.2). Соответственно имеем в этом случае

$$A_2 = A - X Q^T A = (I - X Q^T) A, \quad A_2 X = 0. \quad (32.3)$$

Предположим теперь, что Y состоит из p векторов, линейно независимых с векторами из X и таких, что

$$A Y = [X : Y] \begin{bmatrix} S \\ N \end{bmatrix} = [X : Y] W. \quad (32.4)$$

(Заметим, что Y само по себе не обязано быть инвариантным подпространством A , но из (31.1) и (32.4) следует, что $[X : Y]$ является инвариантным подпространством.) Рассмотрим теперь $A_2(Y - XZ)$, где Z — какая-либо матрица, для которой существует $Y - XZ$. Так как $A_2 X = 0$, то

$$\begin{aligned} A_2(Y - XZ) &= A_2 Y = A Y - X Q^T A Y = [X : Y] W - X Q^T [X : Y] W = \\ &= [X : Y] W - [X : X Q^T Y] W = (Y - X Q^T Y) N. \end{aligned} \quad (32.5)$$

Если мы возьмем $Z = Q^T Y$, то из уравнения (32.5) следует, что $Y - X(Q^T Y)$ есть инвариантное подпространство A_2 . Линейная независимость X и $Y - X(Q^T Y)$ немедленно следует из линейной независимости X и Y . Смысл полученного результата становится более очевидным при рассмотрении практических приложений. Предположим, что A имеет элементарный делитель $(\lambda - \lambda_1)^4$ четвертой степени, и что мы нашли векторы x_1 и x_2 высоты один и два, но не нашли x_3 и x_4 . Если мы произведем исчерпывание, то по (31.4) матрица A_2 имеет два линейных делителя, равных λ , соответствующих x_1 и x_2 , и квадратичный делитель $(\lambda - \lambda_1)^2$, которому соответствуют векторы высоты один и два, отвечающие приведенной форме x_3 и x_4 . На практике, однако, итерации с A дадут, скорее, все пространство, натянутое на x_1, x_2, x_3, x_4 , а не подпространство, натянутое только на x_1 и x_2 .

Практическое приложение

33. Единственное ограничение на матрицу Q заключается в том, что она должна удовлетворять уравнению (32.1). Теоретически это дает некоторую свободу в выборе Q , но на практике Q выбирается таким образом, чтобы минимизировать число вычислений. Если положить $Q^T = [J : K]$ и $X^T = [U^T : V^T]$, где J и U — матрицы m -го порядка, то

очевидно, что простейший выбор Q будет

$$J = U^{-1}, \quad K = 0. \quad (33.1)$$

Если мы обозначим

$$A = \left[\begin{array}{c|c} B & C \\ \hline D & E \end{array} \right], \quad \text{то} \quad A_2 = \left[\begin{array}{c|c} B & C \\ \hline D & E \end{array} \right] - \left[\begin{array}{c} U \\ \hline V \end{array} \right] [U^{-1} \mid 0] \left[\begin{array}{c|c} B & C \\ \hline D & E \end{array} \right], \quad (33.2)$$

что дает

$$A_2 = \left[\begin{array}{c|c} 0 & 0 \\ \hline D - VU^{-1}B & E - VU^{-1}C \end{array} \right], \quad (33.3)$$

так что A_2 имеет равные нулю элементы в строках с 1-й до m -й. Оставшиеся собственные значения совпадают с собственными значениями $E - VU^{-1}C$.

На практике мы не можем взять первые m строк X , а должны использовать схему главных элементов для выбора подходящих m строк. Если это сделано, можно проверить, что матрица $E - VU^{-1}C$ (которая является той частью A_2 , которая действительно используется при дальнейших итерациях) совпадает с соответствующей частью исчерпанной матрицы, полученной по методу § 23.

Однако значительно более удобно перед выполнением исчерпывания привести инвариантное подпространство X к *стандартной трапецевидной форме*. Эта форма получается в результате разложения соответствующей части в произведение треугольных с выбором главного элемента по столбцу. Таким образом, игнорируя выбор главных элементов для удобства обозначений, можем написать

$$X = TR, \quad (33.4)$$

где X , T , и R обычно имеют вид

$$X = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 & 0 \\ \times & 1 & 0 \\ \times & \times & 1 \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}, \quad R = \begin{bmatrix} \times & \times & \times \\ & \times & \times \\ & & \times \end{bmatrix}. \quad (33.5)$$

Здесь T — нижняя единичная трапецевидная матрица, R — верхняя треугольная матрица m -го порядка. Очевидно, что если X — инвариантное подпространство, то это же верно для T , так что мы всюду можем использовать T вместо X . Так как теперь подматрица T , соответствующая U , есть нижняя единичная треугольная матрица m -го порядка, вычисление $E - VU^{-1}C$ довольно просто.

Снова, если A имеет много нулей, мы можем использовать (33.2) в виде

$$A_2 = \left\{ I - \left[\begin{array}{c|c} I & 0 \\ \hline VU^{-1} & 0 \end{array} \right] \right\} A, \quad (33.6)$$

если это более экономно. Из (33.3) мы знаем, что инвариантное подпространство Y матрицы A_2 имеет первые m строк равными нулю, и, следовательно, при итерировании с A_2 можем начинать с начального Y_0 такой

же формы, и все Y_s останутся в этой же форме. Если мы обозначим

$$AY_s = \begin{bmatrix} K_s \\ \dots \\ M_s \end{bmatrix}, \quad (33.7)$$

то A_2Y_s получается из AY_s при помощи вычитания столбцов T , умноженных на множители так, чтобы пропали m первых строк. Это довольно просто, так как T это единичная трапециевидная матрица.

Ступенчатые итерации

34. Одна из трудностей при использовании итераций и исчерпывания состоит в том, что, вообще говоря, мы не знаем на каждом этапе, является ли текущее доминирующее собственное значение вещественным или комплексным и соответствует ли оно линейному или нелинейному элементарному делителю. Можно обойти эту трудность, используя полную систему из n векторов одновременно. Мы будем уверены, что они линейно независимы, если возьмем в качестве n векторов столбцы единичной нижней треугольной матрицы. Если L_s будет на каждом шаге обозначать матрицу, образованную системой n векторов, то процесс будет иметь вид

$$X_{s+1} = AL_s, \quad X_{s+1} = L_{s+1}R_{s+1}, \quad (34.1)$$

где каждая L_s есть единичная нижняя треугольная, а каждая R_s — верхняя треугольная. Если $L_0 = I$, то

$$L_s R_s = X_s = AL_{s-1}, \quad (34.2)$$

$$L_s R_s R_{s-1}^{-1} = AL_{s-1} R_{s-1}^{-1} = AL_{s-2} = A^2 L_{s-2}, \quad (34.3)$$

$$L_s (R_s R_{s-1} \dots R_1) = A^s L_0 = A^s. \quad (34.4)$$

Следовательно, L_s и $R_s R_{s-1} \dots R_1$ суть матрицы, полученные при разложении A^s в произведение треугольных, так что (гл. 8, § 4) L_s равны произведению первых s нижних треугольных матриц, полученных при помощи LR -алгоритма, а R_s совпадают с верхними треугольными матрицами этого алгоритма. Мы знаем, следовательно (гл. 8, § 6), что если, например, модули всех собственных значений различны, то $L_s \rightarrow L$, где L — матрица, полученная при разложении в произведение треугольных матрицы правых собственных векторов A . Не надо предполагать, что это формальное сходство с LR -алгоритмом означает, что этот алгоритм столь же распространен на практике, как и LR -алгоритм.

Успех LR -алгоритма по существу зависит от ряда факторов. Среди них наиболее важными являются инвариантность верхней формы Хессенберга (даже при использовании перестановок), эффективность сдвига и устойчивость процесса исчерпывания. При ступенчатых итерациях, если A — верхняя матрица Хессенберга, и L_0 есть I , то, хотя L_1 имеет только $(n-1)$ ненулевых поддиагональных элементов, число ненулевых элементов в последовательных L_i прогрессивно растет, и L_{n-1} — уже полная нижняя треугольная матрица; с этого момента и AL_i являются полными матрицами.

С другой стороны, если A — нижняя матрица Хессенберга, AL_i также сохраняет эту форму на каждом этапе. Хотя разложение AL_i в произведение треугольных требует лишь $\frac{1}{2} n^2$ умножений, вычисление AL_i тре-

бует уже $\frac{1}{6}n^3$ умножений. Введение перестановок строк при разложении AL_i разрушает преимущества формы Хессенберга, а перестановки столбцов — нет, хотя применение перестановок, как и в LR -алгоритме, не имеет теоретического обоснования. Ускорения сходимости можно достигнуть, применяя $(A - pI)$ вместо A , и p можно менять от итерации к итерации, но отсутствие действительно эффективной техники исчерпывания делает это менее удовлетворительным, чем в LR -алгоритме. Когда малое собственное значение λ_i определено, мы не можем свободно выбирать p , так как это может сделать $(\lambda_i - p)$ доминирующим собственным значением A .

Очевидно, что нам не обязательна полная система n векторов. Мы можем работать с группой из m векторов из единичной нижней трапецевидной формы. Обозначив ее T_s , получаем процесс в виде

$$AT_s = X_{s+1}, \quad X_{s+1} = T_{s+1}R_{s+1}, \quad (34.5)$$

где R_{s+1} — верхние треугольные матрицы m -го порядка. Если доминирующие m собственных значений A имеют различные модули, $T_s \rightarrow T$, где T получена приведением матрицы из m доминирующих собственных векторов к верхней трапецевидной форме. Более общо: если

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_m| > |\lambda_{m+1}|, \quad (34.6)$$

то матрицы T_s не обязательно стремятся к пределу, но подпространства, натянутые на их столбцы, стремятся к инвариантному подпространству (гл. 8, §§ 32, 34). Этот процесс впервые был описан Ф. Л. Бауэром, который назвал его *ступенчатыми итерациями*.

Если $|\lambda_1| \gg |\lambda_2|$, то первый столбец T_s будет сходиться к x_1 в несколько итераций и нет смысла включать x_1 в T_s при вычислении X_{s+1} . Вообще говоря, если первые k векторов T_s уже стабилизировались, нам не надо умножать эти векторы на A при последующих шагах. Можно написать

$$T_s = [T_s^{(1)} : T_s^{(2)}], \quad (34.7)$$

где $T_s^{(1)}$ состоит из первых k векторов, которые уже стабилизировались, а $T_s^{(2)}$ — из остальных $(m - k)$ векторов, которые еще не стабилизировались. Определим X_{s+1} соотношением

$$X_{s+1} = [T_s^{(1)} : AT_s^{(2)}], \quad (34.8)$$

в котором $T_s^{(1)}$ уже в трапецевидной форме, но $AT_s^{(2)}$, вообще говоря, состоит из $(m - k)$ полных векторов. Очевидным образом X_{s+1} затем может быть приведена к трапецевидной форме. Ясно, что процесс, который мы сейчас используем, совпадает с процессом, обсуждавшимся в § 30.

Надо заметить, что не обязательно первые столбцы будут стабилизироваться первыми. Например, если четыре доминирующие собственные значения равны 1, 0,99, 10^{-2} , 10^{-3} , первые два столбца скоро будут порождать подпространство, соответствующее собственным векторам x_1 и x_2 , следовательно, столбцы 3 и 4 будут сходиться. Разделение столбцов 1 и 2 в соответствующие собственные векторы будет происходить значительно медленнее. Если доминирующее собственное значение соответствует нелинейному элементарному делителю, то этот эффект может быть еще более заметным. Численный пример дан в § 41.

Точное определение комплексно сопряженных собственных значений

35. Методы нескольких предыдущих параграфов позволяют определить комплексно сопряженные собственные значения вещественной матрицы A без заметной потери точности, если соответствующие собственные значения имеют малую мнимую часть (конечно, за исключением случая, когда A близка к матрице, имеющей квадратичный элементарный делитель). Опишем процесс, не включающий выбор главных элементов. На каждом шаге мы имеем дело с парой вещественных векторов a_s и b_s в стандартной трапецевидной форме, так что

$$\begin{aligned} a_s^T &= [1, \alpha_s, \times, \dots, \times, \times], \\ b_s^T &= [0, 1, \times, \dots, \times, \times]. \end{aligned} \quad (35.1)$$

Вычислим c_{s+1} и d_{s+1} :

$$c_{s+1} = Aa_s, \quad d_{s+1} = Ab_s, \quad (35.2)$$

и положим

$$\begin{aligned} c_{s+1}^T &= [\beta_{s+1}, \gamma_{s+1}, \times, \dots, \times, \times], \\ d_{s+1}^T &= [\delta_{s+1}, \varepsilon_{s+1}, \times, \dots, \times, \times]. \end{aligned} \quad (35.3)$$

Следующие векторы a_{s+1} и b_{s+1} получаются приведением (c_{s+1}, d_{s+1}) к стандартной трапецевидной форме.

Если доминирующими собственными значениями A является комплексно сопряженная пара $(\xi_1 \pm i\eta_1)$, то асимптотически

$$[c_{s+1} \mid d_{s+1}] = [a_s \mid b_s] M_s, \quad (35.4)$$

где M_s — матрица второго порядка, которая меняется с ростом s , но имеет фиксированные собственные значения $(\xi_1 \pm i\eta_1)$. На каждом шаге определяем матрицу M_s такую, что

$$\begin{bmatrix} 1 & 0 \\ \alpha_s & 1 \end{bmatrix} M_s = \begin{bmatrix} \beta_{s+1} & \delta_{s+1} \\ \gamma_{s+1} & \varepsilon_{s+1} \end{bmatrix}, \quad (35.5)$$

и получаем последовательные приближения к $\xi_1 \pm i\eta_1$ из собственных значений матрицы M_s .

Можно подумать, что если η_1 мало, то мы будем терять точность при вычислении собственных значений M_s , но это будет не так, если $(\xi_1 \pm i\eta_1)$ являются хорошо обусловленными собственными значениями A . В этом случае окажется, что $m_{11}^{(s)}$ и $m_{22}^{(s)}$ будут близки к ξ_1 , а $m_{12}^{(s)}$ и $m_{21}^{(s)}$ будут малы. В самом деле, если η_1 стремится к нулю, то оба $m_{12}^{(s)}$ и $m_{21}^{(s)}$ становятся нулями в силу предположения, что A не имеет квадратичного делителя $(1 - \xi_1)^2$. Векторы a_s и b_s или c_{s+1} и d_{s+1} обязательно порождают инвариантное подпространство точно, но, как мы видели в § 14, определение отдельных собственных векторов является плохо обусловленной проблемой. Если собственный вектор u_s матрицы M_s , соответствующий $\xi_1 + i\eta_1$, имеет вид

$$u_s^T = (\varphi_s, \psi_s), \quad (35.6)$$

то соответствующий собственный вектор A имеет вид $(\varphi_s a_s + \psi_s b_s)$. Естественно, что u_s должен быть нормирован так, чтобы наибольший из $|\varphi_s|$ и $|\psi_s|$ был равен единице.

Для численной устойчивости исключения при приведении $[c_{s+1} : d_{s+1}]$ к трапециевидной форме существенно употребление схемы главных элементов. Заметим, однако, что единичные элементы в a_s и b_s не должны обязательно быть на первых двух местах, более того, их позиции могут меняться с ростом s . Но матрица M_s определяется *четырьмя элементами из c_{s+1} и d_{s+1} , находящимися в тех же позициях, что и единичные элементы в a_s и b_s .*

Этот метод содержит приблизительно вдвое больше вычислений, чем метод § 12, где мы пытались определить $(\xi_i \pm i\eta_i)$ из одной последовательности итерированных векторов.

Очень близкие собственные значения

36. Совершенно аналогичная техника может применяться для точного определения группы близких доминирующих собственных значений, если, конечно, они хорошо обусловлены. Предположим, например, что вещественная матрица A имеет доминирующие собственные значения $\lambda_1, \lambda_1 - \varepsilon_1, \lambda_1 - \varepsilon_2$, где ε_i малы. Пусть мы итерируем, используя трапециевидную систему векторов a_s, b_s, c_s , и пусть

$$[d_{s+1} : e_{s+1} : f_{s+1}] = A [a_s : b_s : c_s]. \quad (36.1)$$

Тогда, так же как в § 35, мы можем определить матрицу M_s третьего порядка по подходящим элементам a_s, b_s, c_s и $d_{s+1}, e_{s+1}, f_{s+1}$. Собственные значения M_s должны стремиться к $\lambda_1, \lambda_1 - \varepsilon_1, \lambda_1 - \varepsilon_2$ соответственно. Если собственные значения A неплохо обусловлены, собственные значения M определяются точно. В самом деле, диагональные элементы M_s будут близки к λ_1 , а недиагональные элементы будут величинами порядка ε_i .

Метод ортогонализации

37. Сейчас рассмотрим другую технику подавления доминирующего собственного вектора (или собственных векторов). Простейший пример ее использования связан с вещественными симметричными матрицами. Предположим, что λ_1 и x_1 определены так, что $Ax_1 = \lambda_1 x_1$. Мы знаем, что остальные собственные векторы ортогональны к x_1 и, следовательно, можем подавлять компоненту по x_1 в любом векторе при помощи ортогонализации его с x_1 . Это дает следующую итерационную процедуру:

$$v_{s+1} = Au_s, \quad w_{s+1} = v_{s+1} - (v_{s+1}^T x_1) x_1, \quad u_{s+1} = w_{s+1} / \max(w_{s+1}), \quad (37.1)$$

где предположено, что $\|x_1\|_2 = 1$. Очевидно, u_s стремится к субдоминирующему собственному вектору или, если A имеет второе собственное значение, равное λ_1 , к собственному вектору, соответствующему λ_1 и ортогональному к x_1 .

Очевидно, что если соотношение $|\lambda_1| \gg |\lambda_2|$ не имеет места, то нет особой необходимости проводить ортогонализацию с x_1 на каждой итерации, хотя если A высокого порядка, работа, связанная с ортогонализацией, сравнительно мала.

Мы можем обобщить этот процесс для нахождения x_{r+1} , когда $x_1, x_2, \dots, x_r$ уже определены. Соответствующие итерации имеют вид

$$v_{s+1} = Au_s, \quad w_{s+1} = v_{s+1} - \sum_{i=1}^r (v_{s+1}^T x_i) x_i, \quad u_{s+1} = w_{s+1} / \max(w_{s+1}). \quad (37.2)$$

Заметим, что с точностью до ошибок округления, результаты этого метода

совпадают с результатами метода Хотеллинга (§ 19), так как из (37.1) имеем

$$\begin{aligned} w_{s+1} &= v_{s+1} - (v_{s+1}^T x_1) x_1 = Au_s - x_1 (u_s^T A x_1) = \\ &= Au_s - \lambda_1 x_1 (u_s^T x_1) = Au_s - \lambda_1 x_1 (x_1^T u_s) = (A - \lambda_1 x_1 x_1^T) u_s. \end{aligned} \quad (37.3)$$

Существует аналогичный процесс, который можно использовать, если A — несимметричная, но он требует вычисления как правого собственного вектора x_1 , так и левого собственного вектора y_1 . Если они нормированы так, что $x_1^T y_1 = 1$, то можно использовать следующую итерационную процедуру:

$$v_{s+1} = Av_s, \quad w_{s+1} = v_{s+1} - (v_{s+1}^T y_1) x_1, \quad u_{s+1} = w_{s+1} / \max(w_{s+1}). \quad (37.4)$$

В силу биортогональности левых и правых собственных векторов компонента по x_1 у w_{s+1} подавлена. Снова мы имеем естественное обобщение на случай, когда определены r собственных векторов.

Аналог ступенчатых итераций, использующий ортогонализацию

38. В методе ступенчатых итераций мы работаем одновременно с несколькими векторами, и их линейная независимость поддерживается приведением их на каждом шаге к стандартной трапецевидной форме. Существует аналогичный метод, в котором линейная независимость векторов поддерживается при помощи процесса ортогонализации Шмидта (гл. 4, § 54) или его аналога. Рассмотрим сначала случай, когда мы одновременно итерируем n векторов. Обозначим матрицу, образованную этими векторами на каждом этапе, через Q_s . Тогда процесс имеет вид

$$AQ_s = V_{s+1}, \quad V_{s+1} = Q_{s+1} R_{s+1}, \quad (38.1)$$

где столбцы Q_{s+1} ортогональны, а R_{s+1} — верхняя треугольная. Следовательно,

$$AQ_s = Q_{s+1} R_{s+1}, \quad (38.2)$$

и если $Q_0 = I$, то имеем

$$A^s = Q_s (R_s R_{s-1} \dots R_1). \quad (38.3)$$

Соотношение (38.3) означает, что Q_s и $(R_s R_{s+1} \dots R_1)$ суть матрицы, получаемые при ортогональном приведении A^s к треугольному виду. Сравнивая с QR -алгоритмом (гл. 8, § 28), мы видим, что Q_s равна произведению первых s ортогональных матриц, определенных при помощи QR -алгоритма. Следовательно, например, если все $|\lambda_i|$ различны, то последовательность Q_s стремится в основном к пределу, который является матрицей, полученной при треугольной ортогонализации матрицы из собственных векторов X . Диагональные элементы R_s стремятся к $|\lambda_i|$. В более общем случае, когда некоторые собственные значения имеют равные модули, соответствующие столбцы Q_s могут не стремиться к пределу, но они обязательно будут порождать соответствующее подпространство. Так же, как и в § 34, мы предупреждаем читателя, что формальная эквивалентность метода этого параграфа с QR -алгоритмом не означает, что он ведет к столь же ценной практической процедуре.

Инвариантность верхней формы Хессенберга и автоматический и устойчивый метод исчерпывания существенны для успеха QR -алгоритма, и мы не имеем никакого сравнимого приема в схеме ортогонализации (ср. с соответствующим обсуждением ступенчатых итераций в § 34).

39. Мы можем использовать для ортогонального приведения к треугольной матрице процесс Хаусхолдера (гл. 4, § 46). В этом случае получаем вместо Q_s транспонированную Q_s^T в виде произведения $P_{n-1}^{(s)} P_{n-2}^{(s)} \dots P_1^{(s)}$ из $(n-1)$ матриц отражения. Процесс будет иметь вид

$$AP_1^{(s)} P_2^{(s)} \dots P_{n-1}^{(s)} = V_{s+1}, \quad P_{n-1}^{(s+1)} \dots P_2^{(s+1)} P_1^{(s+1)} V_{s+1} = R_{s+1}. \quad (39.1)$$

Если все $|\lambda_i|$ различны, диагональные элементы $|R_s|$ сходятся, и на этом этапе можем считать текущую Q_s по $P_i^{(s)}$; нам не надо вычислять Q_s на более ранних этапах. Когда некоторые из $|\lambda_i|$ равны, соответствующие элементы $|R_s|$ могут не сходиться, и в этом случае невозможно различить, когда «предел» уже достигнут в том смысле, что группы столбцов Q_s порождают соответствующие подпространства.

Мы описали случай, когда Q_s имеет полное число столбцов, для того чтобы подчеркнуть связь с QR -алгоритмом. На самом же деле ничто не мешает нам работать с меньшим числом столбцов. Если Q_s состоит из r столбцов, то процесс все равно описывается равенствами (37.1), но теперь R_s это верхние треугольные матрицы r -го порядка. Интересен случай, когда $r = 2$, и доминирующие собственные значения образуют комплексно сопряженную пару $(\xi_1 \pm i\eta_1)$. Столбцы Q_s тогда состоят из двух ортогональных векторов, порождающих соответствующее инвариантное подпространство. Этот метод может быть применен для точного определения $(\xi_1 \pm i\eta_1)$ и инвариантного подпространства даже в том случае, когда η_1 мало. На каждом этапе имеем

$$V_{s+1} = Q_s M_s, \quad (39.2)$$

где M_s — матрица второго порядка. Из ортогональности столбцов Q_s имеем

$$M_s = Q_s^T V_{s+1}. \quad (39.3)$$

Собственные значения M_s стремятся к $(\xi_1 \pm i\eta_1)$, и если η_1 мало, но собственные значения хорошо обусловлены, то $m_{12}^{(s)}$ и $m_{21}^{(s)}$ будут порядка η_1 , так что $(\xi_1 \pm i\eta_1)$ будут точно определены.

Биортогонализация

40. Бауэр (1958) описал обобщение метода § 37, которое он назвал *биортогонализацией*. В нем могут быть одновременно найдены x_i и y_i . На каждом этапе рассматриваются две системы по n векторов, являющиеся столбцами двух матриц X_s и Y_s . Образуются матрицы

$$P_{s+1} = AX_s, \quad Q_{s+1} = A^T Y_s, \quad (40.1)$$

и по ним определяются две матрицы X_{s+1} и Y_{s+1} . Столбцы этих матриц выбираются так, чтобы они были биортогональны. Уравнения, определяющие i -е столбцы X_{s+1} и Y_{s+1} , будут

$$\left. \begin{aligned} x_i^{(s+1)} &= p_i^{(s+1)} - r_{1i} x_1^{(s+1)} - r_{2i} x_2^{(s+1)} - \dots - r_{i-1,i} x_{i-1}^{(s+1)}, \\ y_i^{(s+1)} &= q_i^{(s+1)} - u_{1i} y_1^{(s+1)} - u_{2i} y_2^{(s+1)} - \dots - u_{i-1,i} y_{i-1}^{(s+1)}, \end{aligned} \right\} \quad (40.2)$$

где r_{ki} и u_{ki} выбраны так, что

$$(y_k^{(s+1)})^T x_i^{(s+1)} = 0, \quad (x_k^{(s+1)})^T y_i^{(s+1)} = 0 \quad (k = 1, \dots, i-1). \quad (40.3)$$

Очевидно, мы имеем

$$P_{s+1} = X_{s+1}R_{s+1}, \quad Q_{s+1} = Y_{s+1}U_{s+1}, \quad (40.4)$$

где R_{s+1} и U_{s+1} — единичные верхние треугольные матрицы с элементами r_{ki} и u_{ki} . Из биортогональности имеем

$$Y_{s+1}^T X_{s+1} = D_{s+1}, \quad (40.5)$$

где D_{s+1} — диагональная. Если мы возьмем $X_0 = Y_0 = I$, то получим

$$\left. \begin{aligned} A &= X_1 R_1, & A X_s &= X_{s+1} R_{s+1}, \\ A^T &= Y_1 U_1, & A^T Y_s &= Y_{s+1} U_{s+1}, \end{aligned} \right\} \quad (40.6)$$

что дает

$$A^s = X_s R_s \dots R_2 R_1, \quad (A^T)^s = Y_s U_s \dots U_2 U_1. \quad (40.7)$$

Следовательно,

$$A^{2s} = (Y_s U_s \dots U_2 U_1)^T X_s R_s \dots R_2 R_1 = (U_1^T U_2^T \dots U_s^T) D_s (R_s \dots R_2 R_1), \quad (40.8)$$

что показывает, что единичная нижняя треугольная матрица $U_1^T U_2^T \dots U_s^T$ совпадает с соответствующей матрицей в разложении A^{2s} в произведение двух треугольных. Мы знаем, что в LR -алгоритме $L_1, L_2, \dots, L_{2s}$ это матрица, полученная при разложении A^{2s} в произведение треугольных, и, следовательно,

$$U_s^T = L_{2s-1} L_{2s}. \quad (40.9)$$

Результаты, полученные для LR -алгоритма, поэтому немедленно переносятся на метод биортогонализации. В частности, если $|\lambda_i|$ различны, то $U_1^T U_2^T \dots U_s^T$ стремятся к матрице, полученной при треугольном разложении X , и, следовательно, $U_s \rightarrow I$; аналогично $R_s \rightarrow I$. На практике столбцы X и Y нормируются на каждом этапе обычно так, что $\max(x_i^{(s)}) = \max(y_i^{(s)}) = 1$. Следовательно, если все $|\lambda_i|$ различны, $X_s \rightarrow X$ и $Y_s \rightarrow Y$. Один шаг биортогонализации аналогичен в смысле сходимости двум шагам LR -алгоритма. Если некоторые из $|\lambda_i|$ равны, то подпространства, натянутые на соответствующие столбцы X_s и Y_s , стремятся к соответствующим инвариантным подпространствам. Если A — симметричная, то системы X_s и Y_s совпадают, и метод переходит в метод § 37, так как теперь лишь столбцы X_s ортогонализуются на каждом шаге. Это соответствует тому, что для симметричной матрицы один шаг QR -алгоритма эквивалентен двум шагам LR -алгоритма.

Мы работали с матрицами X_s и Y_s , состоящими из полных систем по n векторов, с целью показать связь с LR - и QR -алгоритмами, но процесс может быть использован, если X_s и Y_s имеют любое число столбцов от 1 до n . Применяются те же соотношения, но R_s и U_s — теперь матрицы r -го порядка. Процесс, вообще говоря, приведет к r левым и правым собственным векторам (или к соответствующим инвариантным подпространствам).

Численный пример

41. С целью обратить внимание на некоторые проблемы, возникающие при использовании методов предыдущих параграфов, приведем простой численный пример. В табл. 2 показывается применение метода ступенчатых итераций (§ 34) и метода ортогонализации (§ 38). Здесь A — мат-

Таблица 2

$$\begin{array}{c}
 A \\
 \left[\begin{array}{ccc} -2,1 & -1,2 & -0,1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right] \quad \begin{array}{l} \lambda_1 = \lambda_2 = -1 \\ \lambda_3 = -0,1 \end{array} \quad \begin{array}{l} \text{Элементарные делители } (\lambda+1)^2, (\lambda+0,1) \end{array} \\
 \\
 \begin{array}{cccccc} x_1 & x_2 & x_3 & y_1 & y_2 & y_3 \\ 1 & -2 & 0,01 & 1,0 & 1 & 1 \\ -1 & 1 & -0,1 & 1,1 & 0,1 & 2 \\ 1 & 0 & 1 & 0,1 & 0 & 1 \end{array}
 \end{array}$$

x_1 и x_2 из инвариантного подпространства A , x_1 — собственный вектор; y_1 и y_2 — инвариантного подпространства A^T , y_1 — собственный вектор.

Ступенчатые итерации

$$\begin{array}{c}
 AI \qquad \qquad \qquad L_1 \qquad \qquad \qquad R_1 \\
 \left[\begin{array}{ccc} -2,1 & -1,2 & -0,1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right] \longrightarrow \left[\begin{array}{ccc} 1 & . & . \\ -0,476 & -0,571 & 1 \\ 0 & 1 & . \end{array} \right] \left[\begin{array}{ccc} -2,1 & -1,2 & -0,1 \\ . & 1 & -0,048 \\ . & . & 0 \end{array} \right]
 \end{array}$$

Заметим, что строки, содержащие главные элементы, находятся в позициях 1, 3, 2 соответственно и остаются там же.

$$\begin{array}{c}
 AL_1 \qquad \qquad \qquad L_2 \qquad \qquad \qquad R_2 \\
 \left[\begin{array}{ccc} -1,529 & 0,585 & -1,200 \\ 1,000 & 0 & 0 \\ -0,476 & -0,571 & 1,000 \end{array} \right] \longrightarrow \left[\begin{array}{ccc} 1 & . & . \\ -0,654 & -0,509 & 1 \\ 0,311 & 1 & . \end{array} \right] \left[\begin{array}{ccc} -1,529 & 0,585 & -1,200 \\ . & . & -0,086 \\ . & -0,753 & 1,373 \end{array} \right]
 \end{array}$$

$$\begin{array}{c}
 AL_2 \qquad \qquad \qquad L_3 \qquad \qquad \qquad R_3 \\
 \left[\begin{array}{ccc} -1,346 & 0,511 & -1,200 \\ 1,000 & 0 & 0 \\ -0,654 & -0,509 & 1,000 \end{array} \right] \longrightarrow \left[\begin{array}{ccc} 1 & . & . \\ -0,743 & -0,502 & 1 \\ 0,486 & 1 & . \end{array} \right] \left[\begin{array}{ccc} -1,346 & 0,511 & -1,200 \\ . & . & -0,097 \\ . & -0,757 & 1,583 \end{array} \right]
 \end{array}$$

$$\begin{array}{c}
 AL_3 \qquad \qquad \qquad L_4 \qquad \qquad \qquad R_4 \\
 \left[\begin{array}{ccc} -1,257 & 0,502 & -1,200 \\ 1,000 & 0 & 0 \\ -0,743 & -0,502 & 1,000 \end{array} \right] \longrightarrow \left[\begin{array}{ccc} 1 & . & . \\ -0,796 & -0,500 & 1 \\ 0,591 & 1 & . \end{array} \right] \left[\begin{array}{ccc} -1,257 & 0,502 & -1,200 \\ . & . & -0,101 \\ . & -0,799 & 1,709 \end{array} \right]
 \end{array}$$

Процесс ортогонализации

$$\begin{array}{c}
 AI \qquad \qquad \qquad Q_1 \qquad \qquad \qquad R_1 \\
 \left[\begin{array}{ccc} -2,1 & -1,2 & -0,1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right] = \left[\begin{array}{ccc} -0,903 & -0,196 & -0,395 \\ 0,430 & -0,414 & -0,789 \\ 0 & -0,889 & -0,474 \end{array} \right] \left[\begin{array}{ccc} 2,326 & 1,084 & 0,090 \\ . & 1,125 & 0,020 \\ . & . & 0,038 \end{array} \right]
 \end{array}$$

$$\begin{array}{c}
 AQ_1 \qquad \qquad \qquad Q_2 \qquad \qquad \qquad R_2 \\
 \left[\begin{array}{ccc} 1,380 & 0,820 & 1,824 \\ -0,903 & -0,196 & -0,395 \\ 0,430 & -0,414 & -0,789 \end{array} \right] = \left[\begin{array}{ccc} 0,810 & 0,425 & 0,402 \\ -0,530 & 0,235 & 0,816 \\ 0,252 & -0,875 & 0,425 \end{array} \right] \left[\begin{array}{ccc} 1,704 & 0,664 & 1,488 \\ . & 0,664 & 1,373 \\ . & 0 & 0,087 \end{array} \right]
 \end{array}$$

$$\begin{array}{c}
 AQ_2 \qquad \qquad \qquad Q_3 \qquad \qquad \qquad R_3 \\
 \left[\begin{array}{ccc} -1,090 & -1,087 & -1,866 \\ 0,810 & 0,425 & 0,402 \\ -0,530 & 0,235 & 0,816 \end{array} \right] = \left[\begin{array}{ccc} -0,748 & -0,523 & -0,439 \\ 0,556 & -0,159 & -0,806 \\ -0,364 & 0,837 & -0,388 \end{array} \right] \left[\begin{array}{ccc} 1,458 & 0,964 & 1,322 \\ . & 0,700 & 1,595 \\ . & . & 0,098 \end{array} \right]
 \end{array}$$

$$\begin{array}{c}
 AQ_3 \\
 \left[\begin{array}{ccc} 0,940 & 1,205 & 1,928 \\ -0,748 & -0,523 & -0,439 \\ 0,556 & -0,159 & -0,806 \end{array} \right]
 \end{array}$$

рица третьего порядка, элементарные делители которой равны $(\lambda+1)^2$ и $(\lambda+0,1)$.

При ступенчатых итерациях для приведения к треугольному виду проводился выбор главного элемента. Позиция главного элемента оста-

валась постоянной на каждом шаге. Это будет всегда, если A имеет вещественные собственные значения, за исключением, быть может, нескольких первых шагов. Можно заметить, что последовательность вторых столбцов матриц L_s быстро стремится к пределу. Это вектор, имеющий нулевую первую компоненту и лежащий в подпространстве, соответствующем квадратичному элементарному делителю. Первые столбцы стремятся к единственному собственному вектору, соответствующему $\lambda = -1$, но, конечно, сходимость медленная (§ 16). Так как A третьего порядка, последний столбец фиксирован как только позиции главных элементов стали постоянными, но $r_{23}^{(s)}$ быстро стремится к λ_3 . Тот факт, что $r_{23}^{(s)}$ сходятся, и вторые столбцы L_s стремятся к пределу значительно быстрее первых, является верным признаком того, что A имеет нелинейный элементарный делитель. (Если мы имеем два равных линейных элементарных делителя, то оба столбца имеют пределы, а если имеем комплексно сопряженную пару собственных значений, то ни один из них не имеет предела.) Можно проверить, что столбцы 1 и 2 порождают инвариантное подпространство. Для этого покажем, что первые два столбца AL_3 и первые два столбца L_3 почти линейно зависимы. В самом деле, по методу наименьших квадратов имеем

$$\begin{bmatrix} -1,257 & 0,502 \\ 1,000 & 0 \\ -0,743 & -0,502 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ -0,743 & -0,502 \\ 0,486 & 1,000 \end{bmatrix} \begin{bmatrix} -1,257 & 0,502 \\ -0,132 & -0,743 \end{bmatrix} = \\ = 10^{-3} \begin{bmatrix} 0 & 0 \\ -0,215 & 0 \\ -0,098 & -2,972 \end{bmatrix}. \quad (41.1)$$

Собственные значения находятся из матрицы второго порядка в соотношении (41.1). Характеристическое уравнение

$$\lambda^2 + 2,000\lambda + 1,000215 = 0, \quad (41.2)$$

но собственные значения плохо обусловлены. Напомним, что это же было верно в примере § 13.

Рассмотрим теперь метод ортогонализации. Мы видим, что третий столбец Q_s стремится к пределу, который равен вектору, ортогональному к подпространству, соответствующему квадратичному элементарному делителю. Этот вектор должен быть равен $[6^{-1/2}, 2,6^{-1/2}, 6^{-1/2}]$, но в нашем примере он не такой точный, как можно было бы ожидать. Это произошло потому, что мы использовали процесс ортогонализации Шмидта вместо использования матриц отражения. В нашем примере точный третий столбец можно легко получить вычислением (единственного) вектора, ортогонального первым двум столбцам. Ни один из первых двух столбцов не сходится быстро. Первый столбец медленно сходится к единственному собственному вектору, соответствующему $\lambda = -1$. В самом деле, с точностью до нормирующего множителя, он совпадает с первым столбцом L_s . Второй столбец так же медленно стремится к вектору из двухмерного подпространства, ортогональному к собственному вектору. Элемент $r_{33}^{(s)}$ быстро стремится к $|\lambda_3|$. Знак λ_3 может быть определен из того обстоятельства, что третий столбец Q_s меняет знак при каждой итерации.

В этом случае у нас нет такого же ясного признака присутствия нелинейного элементарного делителя, как при ступенчатых итерациях, но сходимость третьего вектора показывает, что первые два уже порождают

инвариантное подпространство. Соответствующие собственные значения определяются из первых двух столбцов Q_3 и AQ_3 , которые почти линейно зависимы. По методу наименьших квадратов имеем

$$\begin{bmatrix} 0,940 & 1,205 \\ -0,748 & -0,523 \\ 0,556 & -0,159 \end{bmatrix} - \begin{bmatrix} -0,748 & -0,523 \\ 0,556 & -0,159 \\ -0,364 & 0,837 \end{bmatrix} \begin{bmatrix} -1,320 & -1,134 \\ 0,090 & -0,683 \end{bmatrix} = \\ = 10^{-3} \begin{bmatrix} -0,290 & -0,441 \\ 0,230 & -1,093 \\ 0,190 & -0,105 \end{bmatrix}, \quad (41.3)$$

и характеристическое уравнение матрицы второго порядка есть

$$\lambda^2 + 2,003\lambda + 1,003620 = 0. \quad (41.4)$$

Снова определение собственных значений плохо обусловлено.

В случае наличия вместо нелинейного элементарного делителя комплексно сопряженной пары собственных значений метод ступенчатых итераций имеет ту отрицательную черту, что позиции главных элементов меняются в течение всего процесса итераций. Вектор, соответствующий действительному собственному значению, которое меньше по модулю, чем комплексная пара, не будет теперь стремиться к пределу, так как его нулевой элемент не будет в фиксированном положении, и если мы не будем использовать схему главного элемента, то время от времени возможна сильная численная нестабильность. Ортогональная техника не обладает соответствующим дефектом.

Метод биортогонализации неудовлетворителен при наличии нелинейного элементарного делителя. Первые столбцы X_s и Y_s медленно стремятся к x_1 и y_1 соответственно, причем x_1 и y_1 ортогональны (гл. 1, § 7). Следовательно, из определения r_{ki} и u_{ki} в § 40 биортогонализация остальных столбцов становится все более численно неустойчивой.

Процесс уточнения Ричардсона

42. Ричардсон (1950) описал процесс, который можно использовать для нахождения собственных векторов, если известно хорошее приближение к собственным значениям. Результат основывается на следующем. Если собственные значения известны точно и если

$$u = \sum \alpha_i x_i \quad (42.1)$$

— произвольный вектор, то

$$\prod_{i \neq k} (A - \lambda_i I) u = \alpha_k \prod_{i \neq k} (\lambda_k - \lambda_i) x_k, \quad (42.2)$$

а все остальные компоненты исключены. Следовательно, мы можем по очереди определить все собственные векторы. Если вместо точного значения у нас есть хорошее приближение λ'_i , то предварительное умножение справа на $(A - \lambda'_i I)$ значительно уменьшает компоненту по x_i .

Несмотря на то, что метод кажется весьма сильным, его прямое применение разочаровывает. Рассмотрим определение x_{20} для матрицы, имеющей собственные значения $\lambda_r = 1/r$ ($r = 1, \dots, 20$). Предположим, что $\lambda'_1 = 1 - 10^{-10}$, $\lambda'_r = 1/r$ ($r = 2, \dots, 20$) и что вычисления могут быть выполнены точно. Если мы умножим слева произвольный вектор u на

$\prod_{i \neq 20} (A - \lambda_i I)$, все компоненты исчезнут, кроме компонент по x_1 и по x_{20} , которые станут

$$\alpha_1 10^{-10} \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{3}\right) \dots \left(1 - \frac{1}{19}\right) x_1 = \alpha_1 (10^{-10}/19) x_1 \quad (42.3)$$

и

$$\alpha_{20} \left(\frac{1}{20} - 1 + 10^{-10}\right) \left(\frac{1}{20} - \frac{1}{2}\right) \left(\frac{1}{20} - \frac{1}{3}\right) \dots \left(\frac{1}{20} - \frac{1}{19}\right) x_{20} \approx -\alpha_{20} (20)^{-18} x_{20} \quad (42.4)$$

соответственно. Компонента по x_1 действительно *возрастает* относительно компоненты по x_{20} больше чем в 10^4 раз. Если мы сначала выполним три итерации с $(A - \lambda_1' I)$, то это уменьшит компоненту по x_1 в 10^{30} раз, если выполнять действия точно, но из-за ошибок округления каждая компонента не сможет стать значительно меньше, чем $2^{-t} x_i$. Если собственные значения известны с хорошей точностью, то часто можно определить такие r_i , что $\prod_{i \neq k} (A - \lambda_i' I)^{r_i}$ и дает точный x_k , но, вообще говоря, круг приложимости метода Рундсона довольно узок.

Возведение матрицы в степень

43. Вместо того чтобы составлять последовательность $A_s u_0$ для произвольного u_0 , мы можем прямо строить последовательность степеней матрицы. Это имеет то преимущество, что, имея A^s , мы можем получить A^{2s} при помощи одного матричного умножения. Следовательно, можем построить последовательность $A, A^2, A^4, \dots, A^{2^s}$ и получить

$$A^{2^s} = \sum \lambda_i^{2^s} x_i y_i^T = \lambda_1^{2^s} \{x_1 y_1^T + \sum_2^n (\lambda_i/\lambda_1)^{2^s} x_i y_i^T\}. \quad (43.1)$$

Если все $|\lambda_i|$ различны, то в степени матрицы доминирует член $\lambda_1^{2^s} x_1 y_1^T$, так что все строки становятся параллельными y_1^T и все столбцы — параллельными x_1 . Скорость сходимости определяется скоростью, с какой $(\lambda_2/\lambda_1)^{2^s}$ стремится к нулю. Следовательно, сравнительно небольшое число итераций требуется даже тогда, когда $|\lambda_1|$ и $|\lambda_2|$ довольно плохо отделены.

Вообще, последовательные матрицы A^{2^s} будут либо увеличиваться, либо уменьшаться по величине, и даже при вычислениях с плавающей запятой возможны переполнение или обращение в машинный нуль. Этого можно легко избежать, умножая каждую матрицу на степень двух так, чтобы наибольший по модулю элемент имел нулевой порядок. Это не вносит дополнительных ошибок округления. Число итераций, необходимых для исчезновения компоненты по $x_2 y_2^T$ с принятой точностью, при работе со степенями матрицы и при простом степенном методе соответственно определяется из неравенств

$$|\lambda_2/\lambda_1|^{2^s} < 2^{-t} \quad \text{и} \quad |\lambda_2/\lambda_1|^s < 2^{-t} \quad (43.2)$$

Если A не содержит сколько-либо значительного числа нулей, то возведение матрицы в квадрат требует в n раз больше работы, чем простая итерация. Следовательно, возводить матрицы в квадрат выгодно, если

$$t \log 2 / \log |\lambda_1/\lambda_2| > n \log_2 \{t \log 2 / \log |\lambda_1/\lambda_2|\}. \quad (43.3)$$

При данном разделении собственных значений возведение в квадрат менее выгодно при больших n . Наоборот, при фиксированном n метод возведения в квадрат более выгоден, если собственные значения λ_1 и λ_2 плохо отделены. Если A — симметричная, то все степени A тоже симметричные, поэтому мы имеем выигрыш от симметрии при возведении в степень, а в простом итерационном процессе выигрыша от симметрии нет.

Например, в случае $|\lambda_1/\lambda_2| = 1-2^{-10}$ и $t = 40$ требуется 15 шагов возведения в квадрат или около 28 000 шагов простого степенного процесса для уничтожения компонент по x_2 с рабочей точностью. Следовательно, возведение в квадрат более выгодно для матриц приблизительно до 2000-го порядка, причем для матрицы 100-го порядка возведение в квадрат выгоднее почти в 20 раз.

Однако такое сравнение методов несколько сомнительно по двум причинам.

(i) Вряд ли мы станем проводить 28 000 шагов простого степенного процесса без применения каких-либо методов ускорения сходимости, а методы ускорения сходимости трудно применимы в процессе возведения матриц в квадрат.

(ii) Большинство матриц порядка более 50, возникающих на практике, обычно содержит много нулей. Это свойство разрушается при возведении в степень, и, следовательно, наши оценки количества необходимой работы в двух процессах не реальны.

Численная стабильность

44. В случае определения доминирующего собственного значения и собственного вектора метод возведения матрицы в степень весьма точен, особенно если скалярные произведения накапливаются. Интересно посмотреть, что случится, если собственному значению или собственным значениям с максимальным модулем соответствует более одного собственного вектора. Если мы положим

$$A = [a_1 | a_2 | \dots | a_n], \quad \text{то} \quad A^{2^s} = A^{2^s-1} [a_1 | a_2 | \dots | a_n]. \quad (44.1)$$

Каждый вектор a_k можно выразить через векторы x_i . Следовательно, если, например, мы имеем вещественное доминирующее собственное значение с r линейно независимыми собственными векторами, то все эти собственные векторы должны быть представлены в столбцах A^{2^s} . С другой стороны, если A имеет доминирующие собственные значения λ_1 и $-\lambda_1$, каждый столбец A^{2^s} будет состоять из вектора, лежащего в подпространстве, порожденном двумя соответствующими собственными векторами. В любом случае собственные значения A получают наилучшим образом при применении простого степенного метода к линейно независимым векторам-столбцам A^{2^s} .

Так как при образовании высоких степеней A разделение собственных значений ухудшается, на первый взгляд кажется привлекательным комбинировать возведение матриц в степень с простым степенным методом и исчерпыванием. При таком комбинировании мы можем вычислить $A^{2^r} = B_r$ для сравнительно небольших r , и затем осуществить простые итерации, используя B_r . Недостатком такого способа является то, что собственные значения с меньшим модулем сравнительно слабо представлены

в B_r . В самом деле, имеем

$$A^{2^s} = \lambda_1^{2^s} \{x_1 y_1^T + \sum_2^n (\lambda_i / \lambda_1)^{2^s} x_i y_i^T + O(2^{-t})\}, \quad (44.2)$$

где $O(2^{-t})$ используется для обозначения ошибок округления. Если $|\lambda_k| < \frac{1}{2} |\lambda_1|$, то член $x_k y_k^T$ меньше ошибок округления уже при $s = 6$, и из B_6 нельзя получить ни одного знака для x_k и λ_k . Однако комбинирование r шагов возведения матрицы в квадрат и затем простых итераций часто является наиболее экономичным способом нахождения доминирующего собственного значения. В случае, когда несколько других собственных значений близки к доминирующему собственному значению, они могут быть довольно точно найдены из B_r использованием простых итераций и исчерпывания. Так, например, если собственные значения матрицы в убывающем порядке равны

$$1,00, \quad 0,99, \quad 0,98, \quad 0,50, \quad \dots, \quad 0,00,$$

то при шести шагах возведения матрицы в квадрат все собственные значения, за исключением первых трех, фактически пропадут. Первые станут $1, (0,99)^{64}, (0,98)^{64}$ соответственно. Три соответствующих собственных вектора полностью представлены в B_r и могут быть получены простыми итерациями и исчерпыванием. Так как собственные значения теперь сравнительно хорошо разделены, требуется лишь несколько итераций.

Использование полиномов Чебышева

45. Предположим, мы знаем, что все собственные значения от λ_2 до λ_n лежат в некотором интервале (a, b) и что λ_1 лежит вне этого интервала. Для эффективной итерационной схемы мы хотим определить функцию $f(A)$ матрицы A так, чтобы $|f(\lambda_1)|$ было как можно больше по сравнению с максимально возможными значениями $|f(\lambda)|$ в интервале (a, b) . Если мы ограничимся полиномами, то решение этой проблемы дают полиномы Чебышева.

Применение преобразования

$$x = [2\lambda - (a + b)] / (b - a) \quad (45.1)$$

переводит интервал (a, b) в $(-1, 1)$, и полином $f_n(\lambda)$ степени n , удовлетворяющий нашим требованиям, равен

$$f_n(\lambda) = T_n(x) = \cos(n \arccos x). \quad (45.2)$$

Следовательно, если мы определим B соотношением

$$B = [2A - (a + b)I] / (b - a), \quad (45.3)$$

то наиболее эффективным полиномом степени s будет $T_s(B)$. Полиномы Чебышева удовлетворяют рекуррентным соотношениям

$$T_{s+1}(B) = 2BT_s(B) - T_{s-1}(B), \quad (45.4)$$

и, следовательно, начиная с произвольного вектора v_0 , мы можем образовывать векторы $v_s = T_s(B)v_0$ для $s = 1, 2, \dots$ из соотношений

$$T_{s+1}(B)v_0 = 2BT_s(B)v_0 - T_{s-1}(B)v_0, \quad (45.5)$$

$$v_{s+1} = 2Bv_s - v_{s-1}. \quad (45.6)$$

Для каждого вектора нужно лишь одно умножение, т. е. не больше, чем при простых итерациях. Очевидно, что подобным образом может быть найден и собственный вектор, соответствующий λ_n .

Существует обширная литература об использовании ортогональных полиномов для вычисления собственных векторов. Такие методы широко используются для вычисления наименьшего собственного значения больших матриц, содержащих большое количество нулей, которые получаются из конечноразностных приближений к уравнениям в частных производных. Вообще говоря, успех весьма существенно зависит от деталей практической процедуры. Сошлемся на работы Штифеля (1958) и Энгели и др. (1959).

Общая оценка методов, основанных на прямых итерациях

46. Сейчас можно оценить методы, основанные на итерациях с A или $(A - pI)$. Основная слабость этих методов состоит в ограниченной скорости сходимости. Их главное достоинство в простоте и в том, что они сохраняют все преимущества матриц, имеющих много нулевых элементов, при условии, что не производится явное исчерпывание.

Мне кажется, что наиболее частое поле их приложения — это определение небольшого числа алгебраически наибольших или наименьших собственных значений и соответствующих собственных векторов больших матриц, имеющих много нулевых элементов. Для очень больших редких матриц они обычно являются единственными возможными методами, так как методы, использующие приведение матриц к некоторому специальному виду, обычно нарушают присутствие нулей и требуют непозволительно много времени. Если ненулевые элементы матрицы, имеющей много нулей, определяются каким-либо систематическим способом, часто можно избежать запоминания матрицы и по мере надобности строить ее элементы в течение каждой итерации. Если нужно определить более, чем один собственный вектор, то исчерпывание надо производить не явно, а так, как в § 30.

Методы, в которых используется несколько векторов (или даже полная система n векторов), имеют большое теоретическое значение из-за их связи с LR - и QR -алгоритмами. На практике редко встречаются ситуации, когда их применение может быть рекомендовано, хотя если известно (или ожидается), что имеются нелинейные элементарные делители, они используются для вычисления соответствующих инвариантных подпространств с достаточной точностью.

Ожидалось, что они могли бы быть использованы для уточнения систем собственных значений и собственных векторов, определенных методами, связанными с преобразованиями подобия A , и вносящими, следовательно, ошибки округления. Однако, как мы уже указывали, стабильность лучших методов, использующих преобразования подобия, замечательна, в то время как большая стабильность методов, непосредственно использующих A , не так заметна, как это можно было бы ожидать. Я считаю, что метод, который мы обсудим в § 54, более пригоден для уточнения полной системы собственных векторов и собственных значений, чем ступенчатые итерации и родственные методы. Тем не менее я знаю по собственному опыту, что небольшие модификации алгоритма иногда существенно улучшают его эффективность; возможно, что это справедливо и для ступенчатых итераций, методов ортогонализации и биортогонализации.

Обратные итерации

47. Мы уже отмечали, что LR - и QR -алгоритмы связаны со степенным методом. Поэтому кажется странным, что, в то время как в этих алгоритмах достигается большая скорость сходимости, основным недостатком алгоритмов, основанных на степенном методе, является малая скорость сходимости.

Если для сходимости к λ_n используется сдвиг p , скорость сходимости LR - и QR -алгоритмов определяется скоростью, с которой $[(\lambda_n - p)/(\lambda_{n-1} - p)]^s$ стремится к нулю. В этих процессах p выбрано приближением к λ_n . Мы можем получить столь же высокую скорость сходимости степенного метода, если будем итерировать с $(A - pI)^{-1}$, а не $(A - pI)$. С точностью до ошибок округления процесс имеет следующий вид:

$$v_{s+1} = (A - pI)^{-1} u_s, \quad u_{s+1} = v_{s+1} / \max(v_{s+1}). \quad (47.1)$$

Если

$$u_0 = \sum \alpha_i x_i, \quad (47.2)$$

то с точностью до нормирующего множителя имеем

$$u_s = \sum \alpha_i (\lambda_i - p)^{-s} x_i. \quad (47.3)$$

Следовательно, если $|\lambda_k - p| \ll |\lambda_i - p|$ ($i \neq k$), компоненты по x_i ($i \neq k$) в u_s быстро уменьшаются с ростом s . Мы уже описывали этот процесс в связи с вычислением собственных векторов симметричной трехдиагональной матрицы, но ввиду его большой практической ценности мы исследуем его в этом более общем случае.

Анализ ошибок при обратных итерациях

48. Аргументы предыдущего параграфа показывают, что, вообще говоря, чем ближе p к собственному значению λ_k , тем быстрее u_s сходится к x_k . Однако при стремлении p к λ_k матрица $(A - pI)$ становится все более и более плохо обусловленной и, наконец, при $p = \lambda_k$ она вырождается. Важно понять влияние этого обстоятельства на наш итерационный процесс.

На практике матрицу $(A - pI)$ разлагают в произведение LU двух треугольных, используя выбор главного элемента по столбцу, причем записываются элементы L и U и позиции главных элементов (гл. 4, § 39). Это означает, что с точностью до ошибок округления,

$$P(A - pI) = LU, \quad (48.1)$$

где P — некоторая матрица перестановок. С сохраненной информацией мы можем решить уравнение $(A - pI)x = b$, решая две треугольные системы уравнений

$$Ly = Pb \quad \text{и} \quad Ux = y. \quad (48.2)$$

Перестановки здесь играют малую роль, за исключением обеспечения численной стабильности. Эффект ошибок округления таков, что мы имеем

$$P(A - pI + F) = LU, \quad (48.3)$$

и вычисленный вектор x удовлетворяет соотношениям

$$P(A - pI + F + G_b)x = Pb \quad \text{или} \quad (A - pI + F + G_b)x = b, \quad (48.4)$$

где F и G_b суть матрицы с малыми элементами, зависящими от способа округления (гл. 4, § 63). Очевидно, что матрица F не зависит от b , а G_b может зависеть.

Следовательно, зная разложение в произведение треугольных, мы можем вычислить последовательности u_s и v_s вида

$$(A - pI + F + G_{u_s})v_{s+1} = u_s, \quad u_{s+1} = v_{s+1}/\|v_{s+1}\|_2, \quad (48.5)$$

где для удобства анализа мы нормировали u_{s+1} при помощи 2-нормы. Если бы матрицы G_{u_s} были нулевыми, то мы бы итерировали точно с $(A - pI + F)^{-1}$, и, следовательно, u_s стремились бы к собственному вектору $(A + F)$, соответствующему ближайшему к p собственному значению. Точность вычисленного вектора была бы высока, если бы $\|F\|/\|A\|$ было мало и *собственный вектор* был неплохо обусловлен. Как мы уже замечали, вычисленные решения треугольной системы уравнений обычно исключительно точны (гл. 4, §§ 61, 62), так что на практике v_{s+1} обычно очень близко к решению уравнения $(A - pI + F)x = u_s$ на каждом шаге, и итерированные векторы действительно будут сходиться к собственным векторам $(A + F)$.

49. Однако если мы отвлечемся от этих весьма специальных черт треугольных матриц, мы все равно можем показать, что обратные итерации дают хороший результат, даже если p очень близко к λ_k или в точности равно ему, если только x_k неплохо обусловлен. Положим

$$p = \lambda_k + \varepsilon_1, \quad F + G_{u_s} = H, \quad \|H\|_2 = \varepsilon_2 \quad (49.1)$$

и предположим без существенной потери общности, что $\|A\|_2 = 1$. Следовательно, можем написать

$$(A - pI + H)v_{s+1} = u_s. \quad (49.2)$$

Покажем, что для установления нашего результата достаточно убедиться, что v_{s+1} велик. Из (49.2) имеем:

$$(A - \lambda_k I - \varepsilon_1 I + H)v_{s+1}/\|v_{s+1}\|_2 = u_s/\|v_{s+1}\|_2, \quad (49.3)$$

$$(A - \lambda_k I)u_{s+1} = (\varepsilon_1 I - H)u_{s+1} + u_s/\|v_{s+1}\|_2 = \eta, \quad (49.4)$$

где

$$\|\eta\|_2 < (|\varepsilon_1| + \varepsilon_2) + 1/\|v_{s+1}\|_2. \quad (49.5)$$

Это показывает, что если $\|v_{s+1}\|_2$ велика, а ε_1 и ε_2 малы, то u_{s+1} и λ_k дают малую невязку, и следовательно, если x_k неплохо обусловлен, u_{s+1} — хороший собственный вектор. Действительно, положив

$$u_s = \sum \alpha_i x_i, \quad v_{s+1} = \sum \beta_i x_i \quad (49.6)$$

и предположив, что $\|x_i\|_2 = \|y_i\|_2 = 1$, получаем из уравнения (49.2) умножением на y_k^T слева

$$\varepsilon_1 \beta_k + y_k^T H v_{s+1} / s_k = \alpha_k. \quad (49.7)$$

Следовательно,

$$|\alpha_k| \leq |\varepsilon_1| |\beta_k| + \varepsilon_2 \|v_{s+1}\|_2 / s_k \leq |\varepsilon_1| \|v_{s+1}\|_2 / s_k + \varepsilon_2 \|v_{s+1}\|_2 / s_k, \quad (49.8)$$

что дает

$$\|v_{s+1}\|_2 \geq |\alpha_k s_k| / (|\varepsilon_1| + \varepsilon_2). \quad (49.9)$$

Подставив в (49.5), получим

$$\|\eta\|_2 < (|\varepsilon_1| + \varepsilon_2) (1 + 1/|\alpha_k s_k|). \quad (49.10)$$

Следовательно, предполагая, что ни α_k , ни s_k не малы, получаем, что вектор невязки η мал.

Общие комментарии к анализу

50. Обратные итерации имеют следующие потенциальные слабости:

(i) Так как L и U получены при помощи выбора главного элемента по столбцу, существует некоторая опасность того, что элементы U могут быть значительно большими, чем элементы A , и в этом случае ε_2 не будет малыми. Как мы заметили в гл. 4, § 57, это сравнительно слабая опасность. Ее можно обойти, используя полный выбор главного элемента, или вообще устранить, выполняя приведение $(A - pI)$ к треугольной при помощи матриц отражения. Я придерживаюсь мнения, что эта опасность незначительная (см. также §§ 54, 56).

(ii) Наш анализ не показал, что компонента $\beta_k x_k$ на самом деле велика, он показал лишь, что велика $\|v_{s+1}\|$, если u_s имеет составляющую по x_k . Заметим, что мы не можем показать, что β_k велик, если не сделаем дополнительных предположений относительно расположения собственных значений. Если есть несколько собственных значений, очень близких к p , то большая величина $\|v_{s+1}\|$ может быть результатом наличия большой компоненты по любому соответствующему собственному вектору. Это неизбежная слабость; если существуют другие собственные значения, близкие к p , это означает, что x_k плохо обусловлен, и мы должны ожидать некоторых неприятностей из-за этого. Но мы можем показать, что v_{s+1} не может содержать больших компонент по любому x_i , для которого $(\lambda_i - p)$ не мала. Действительно, приравнявая компоненты по x_i в обеих частях (49.2), получаем

$$(\lambda_i - p)\beta_i + y_i^T H v_{s+1}/s_i = \alpha_i, \quad (50.1)$$

$$|\beta_i| \leq \{\varepsilon_2 \|v_{s+1}\|_2 / |s_i| + |\alpha_i|\} |\lambda_i - p|. \quad (50.2)$$

Если λ_k — единственное собственное значение, близкое к p , то большая величина v_{s+1} обязательно означает, что компонента по x_k велика.

(iii) Если исходный вектор u_0 имеет очень малые компоненты по всем x_i , соответствующим λ_i , близким к p , то ошибки округления могут помешать этим компонентам увеличиваться и мы можем никогда не достигнуть u_s , у которого эти компоненты будут существенными. Такая ситуация возможна, но она крайне мало вероятна. По моему опыту, если p близко к λ_k , то начальный вектор, имеющий чрезвычайно малую компоненту по x_k , обычно имеет существенную компоненту уже после одной итерации.

Дальнейшее уточнение собственных векторов

51. Я считаю, что обратные итерации являются самым сильным и точным методом из тех, которые я использовал для вычисления собственных векторов. Для иллюстрации его силы мы опишем, как его можно использовать для получения результатов, соответствующих результатам точ-

ных вычислений. Из нашего анализа ясно, что, работая с фиксированным разложением $(A - pI)$ в произведение двух треугольных, мы не можем ожидать результатов лучших, чем точный собственный вектор матрицы $(A + F)$ из уравнения (48.3). Мы знаем, что матрица F весьма мала, но если существуют другие собственные значения, близкие к λ_k , то вычисленный собственный вектор $(A + F)$ обязательно будет иметь составляющие по соответствующим собственным векторам, которыми нельзя пренебречь. Например, если $\|A\| = O(1)$, $\lambda_1 = 1$, $\lambda_2 = 1 - 2^{-m}$, а другие собственные значения хорошо отделены от этих двух, вычисленный собственный вектор u будет в основном иметь вид

$$x_1 + \varepsilon x_2 + \sum_3^n \varepsilon_i x_i,$$

где ε может быть порядка $2^m \|F\|/|s_2|$, а ε_i порядка $\|F\|/|s_i|$ (гл. 2, § 10).

Интересно обсудить вопрос, можем ли мы получить более точный собственный вектор, чем u , используя то же самое разложение в произведение треугольных и ту же точность вычислений. Рассмотрим точное решение уравнения

$$(A - pI)x = u. \quad (51.1)$$

Очевидно, мы имеем

$$x = \frac{x_1}{p-1} + \frac{\varepsilon x_2}{p-1-2^{-m}} + \sum_3^n \frac{\varepsilon_i x_i}{p-\lambda_i}, \quad (51.2)$$

и следовательно, если p ближе к единице, чем к $1 - 2^{-m}$, то нормированный x будет иметь значительно меньшую компоненту по x_2 , чем u . В гл. 4, §§ 68, 69 мы обсуждали проблему точного решения системы уравнений и показали, что (51.1) может быть решена достаточно точно, если накапливать скалярные произведения и если $(A - pI)$ не слишком плохо обусловлена.

Таким образом, мы предлагаем схему, состоящую из двух существенно различных стадий. На первой стадии мы вычисляем последовательность u_s вида

$$(A - pI)v_{s+1} = u_s, \quad u_{s+1} = v_{s+1}/\max(v_{s+1}). \quad (51.3)$$

Продолжаем такие итерации до тех пор, пока u_s не перестанут улучшаться. Переименовав последний u_s в u , решаем уравнение $(A - pI)x = u$, используя то же самое разложение в произведение треугольных. Это тоже итерационный процесс, имеющий много общего с итерациями на первой стадии. Однако цели их существенно различны.

52. Интересно рассмотреть, как выбрать p так, чтобы получить наибольшую эффективность. Чем ближе p к λ_1 , тем быстрее на первом этапе u_s достигнут предельной точности. Действительно, если мы возьмем $p = \lambda_1 + O(2^{-t})$, то предельная точность почти всегда достигается в две итерации. Плохая обусловленность $(A - pI)$ на этой стадии несущественна. Возвращаясь теперь ко второй стадии, видим, что если мы в состоянии решить $(A - pI)x = u$ точно, то снова выгодно брать p как можно ближе к λ_1 , так как при этом относительная величина компоненты по x_2 в x будет наименьшей. С другой стороны, чем ближе p к λ_1 , тем более плохо обусловленной становится $(A - pI)$ и тем медленнее x сходится к точному решению $(A - pI)x = u$. Наконец, $(A - pI)$ может

стать настолько плохо обусловленной, что хорошее решение вообще не может быть получено, если не работать с большей точностью.

Для того чтобы избежать осложнений, вызванных плохой обусловленностью собственных значений, предположим, что A — симметричная, так что все ее собственные значения хорошо обусловлены. Для удобства предположим, что $\|A\|_2 = 1$, и с целью иллюстрации положим $\lambda_1 = 1$, $\lambda_2 = 1 - 2^{-10}$, $|\lambda_i| < 0,5$ ($i > 2$), $t = 48$. Величина матрицы F зависит от способа округления, и мы предположим, что $\|F\|_2 = n^{1/2} 2^{-48}$. Если мы обозначим собственные значения и собственные векторы $(A + F)$ через λ'_i и x'_i , то получим

$$|\lambda'_i - \lambda_i| < n^{1/2} 2^{-48}, \quad x'_1 = x_1 + \sum_2^n \varepsilon_i x_i, \quad (52.1)$$

где

$$|\varepsilon_2| < n^{1/2} 2^{-48} / 2^{-10} = n^{1/2} 2^{-38}, \quad |\varepsilon_i| < n^{1/2} 2^{-48} / 0,5 = n^{1/2} 2^{-47} \quad (i > 2). \quad (52.2)$$

Предполагая, что p заметно ближе к λ_1 , чем к λ_2 , на первой стадии имеем сходимость к x'_1 , причем меняется лишь *скорость* сходимости, но не предельная точность. Компонента по x'_2 будет уменьшаться относительно компоненты по x'_1 в $(\lambda'_1 - p)/(\lambda'_2 - p)$ раз за итерацию, а другие компоненты уменьшаются еще быстрее. Оценим эффективность процесса для двух различных значений p .

(i) $p = 1 - 2^{-48}$. На первой стадии компонента по x'_2 уменьшается по сравнению с компонентой по x'_1 на множитель, меньший чем

$$(n^{1/2} + 1) 2^{-48} / [2^{-10} - (n^{1/2} + 1) 2^{-48}].$$

Если n не очень велико, то двух итераций почти наверняка достаточно. На второй стадии $(A - pI)$ столь плохо обусловлена, что мы вообще не можем решить $(A - pI)x = u$.

(ii) $p = 1 - 2^{-30}$. На первой стадии компонента по x'_2 уменьшается по сравнению с компонентой по x'_1 на множитель, меньший чем

$$(2^{-30} + n^{1/2} 2^{-48}) / (2^{-10} - 2^{-30} - n^{1/2} 2^{-48}).$$

Следовательно, мы выигрываем приблизительно 20 двоичных знаков за итерацию, и трех итераций почти наверняка будет достаточно. На второй стадии относительная ошибка в последовательных решениях $(A - pI)x = u$ уменьшается приблизительно в $n^{1/2} 2^{30} 2^{-48} = n^{1/2} 2^{-18}$ раз за шаг. Например, при $n = 64$ мы безусловно получим правильно округленное решение этого уравнения за четыре итерации. Точное решение будет

$$x = \frac{x_1}{\lambda_1 - p} + \sum_2^n \frac{\varepsilon_i x_i}{\lambda_i - p}, \quad (52.3)$$

и, следовательно, после нормирования

$$x = x_1 + \sum_2^n \eta_i x_i; \quad |\eta_2| < n^{1/2} 2^{-38} 2^{-20}, \quad |\eta_i| < n^{1/2} 2^{-47} 2^{-30} \quad (i > 2). \quad (52.4)$$

Это показывает, что вычисленный x есть почти правильно округленный x_1 , и плохая обусловленность собственного вектора и ошибки при разложении в произведение треугольных полностью обойдены.

Если λ_1 и λ_2 еще более близки, чем в примере, то вектор x , полученный этим способом, может не быть точным x_1 . В этом случае вторая стадия процесса может быть повторена для получения точного решения

$(A - pI) y = x$, и так далее. Если $|\lambda_1 - \lambda_2| < 2n^{\frac{1}{2}} 2^{-t}$, то собственный вектор x_1 не может быть отделен, если не работать с более высокой точностью.

На основе этой техники было разработано несколько программ для DEUCE. Наиболее усложненная из них была предназначена для определения собственных векторов комплексных матриц с двойной точностью, начиная с приближенных собственных значений. При этом использовалось разложение в произведение треугольных лишь с обычной точностью! Используя технику, описанную в гл. 4, § 69, возможно определить вектор с двойной точностью, потому что, как установлено в пункте (i) настоящего параграфа, мы можем использовать правую часть в уравнении $(A - pI) x = b$ с двойной точностью без применения двойного количества знаков. Программа для DEUCE была весьма замысловатой, и на вычислительных машинах с большой оперативной памятью, наверно, было бы более экономно использовать обычные вычисления с двойной точностью. Однако программа существенно помогла понять природу обратных итераций. Она также дала очень надежный критерий точности вычисленных векторов. Интересно, что эти надежные оценки могут быть получены для λ_i и x_i с использованием лишь грубых оценок остальных собственных значений.

Нелинейные элементарные делители

53. Если A имеет нелинейные элементарные делители, это же верно и для $(A - pI)^{-1}$ при всех значениях p , а мы видели в § 16, что сходимость к доминирующему собственному вектору «медленная», когда он соответствует нелинейному делителю. Может показаться на первый взгляд, что применение обратных итераций в этом случае не дает выигрыша, но это не так.

Рассмотрим простую жорданову подматрицу

$$A = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}. \quad (53.1)$$

Если мы возьмем $p = a - \varepsilon$, то получим

$$(A - pI)^{-1} = \begin{bmatrix} \varepsilon^{-1} & -\varepsilon^{-2} \\ 0 & \varepsilon^{-1} \end{bmatrix}, \quad (53.2)$$

и с точностью до постоянного множителя обратные итерации эквивалентны прямым итерациям с матрицей

$$\begin{bmatrix} \varepsilon & -1 \\ 0 & \varepsilon \end{bmatrix}. \quad (53.3)$$

Если ϵ мало, то, начиная с произвольного вектора, мы немедленно получаем хорошее приближение к единственному собственному вектору в одну итерацию, хотя известно, что последующие итерации не дают существенного исправления. (Это соответствует в анализе § 15 случаю, когда a мало.)

Обратные итерации с матрицами Хессенберга

54. Мы обсудили обратные итерации в связи с полной матрицей A , но на практике более удобно найти собственные векторы некоторой подготовленной матрицы B , которая является преобразованием подобия исходной матрицы, и по ним определить собственные векторы A . Сначала рассмотрим матрицы в верхней форме Хессенберга, так как существуют весьма устойчивые методы приведения матрицы к этой форме.

Разложение матрицы Хессенберга в произведение треугольных с использованием выбора главного элемента по столбцу было обсуждено в гл. 4, § 33. Оно особенно просто, так как, говоря в терминах метода Гаусса, существует только один отличный от нуля множитель на каждом шаге приведения. Всего требуется $\frac{1}{2}n^2$ умножений по сравнению с $\frac{1}{3}n^3$ для полной матрицы. Так как L содержит лишь $(n-1)$ поддиагональных элементов, решение $Ly = Pb$ (см. уравнение (48.2)) требует n умножений. Поэтому в каждой итерации требуется приблизительно $\frac{1}{2}n^2$ умножений сравнительно с n^3 для полной матрицы.

Как в гл. 5, § 54, мы не будем предполагать начальный вектор u произвольным, а выберем его так, чтобы было

$$Le = Pu, \quad (54.1)$$

где e — вектор из единичных элементов. Следовательно, v_1 найдется как решение $Uv_1 = e$. Если p — очень хорошее приближение к собственному значению, то обычно достаточно двух итераций. Если p — точное собственное значение и разложение в произведение треугольных выполнено точно, то один из диагональных элементов U (последний, если среди поддиагональных элементов H нет нулей) должен быть нулем. На практике, однако, может случиться, что ни один из вычисленных диагональных элементов не будет мал. Если диагональный элемент U равен нулю, его можно заменить на $2^{-l} \|H\|_\infty$, но значительно более экономно использовать способ, описанный в § 55.

Мы упоминали здесь, что вектор, полученный по методу Химана (гл. 7, § 11), является точным собственным вектором, если величина z есть точное собственное значение и вычисления выполнены точно. Однако, несмотря на то, что метод Химана определяет собственные значения очень устойчиво, определение собственного вектора часто крайне не устойчиво, и метод Химана нельзя использовать для этой цели. На самом деле, если H — трехдиагональная, вектор, полученный по методу Химана, в точности совпадает с вектором, полученным по методу гл. 5, § 48, а мы уже показали, почему он неустойчивый. При помощи совершенно такого же рассуждения можно показать, что вообще метод Химана дает плохое приближение к x_r , когда нормированный левый собственный вектор y_r имеет малую первую компоненту. Это верно даже в том случае, когда метод Химана выполняется точно с использованием значения λ_r , верного с рабочей точностью.

Вырожденные случаи

55. Если один из поддиагональных элементов $h_{i+1,i}$ равен нулю, то матрицу H можно записать в виде

$$H = \begin{bmatrix} H_1 & B \\ 0 & H_2 \end{bmatrix}, \quad (55.1)$$

где H_1 и H_2 в форме Хессенберга. Собственные значения H совпадают с собственными значениями H_1 и H_2 . Собственные векторы, соответствующие собственным значениям, принадлежащим H_1 , даются соответствующими собственными векторами H_1 , дополненными нулями в последних $(n - i)$ позициях, и следовательно, для их вычисления требуется меньше работы. Собственный вектор, соответствующий собственному значению H_2 , не имеет, вообще говоря, специфической формы, и наиболее экономным способом его получения будет обычный метод обратной итерации.

Если H имеет собственное значение кратности r с r линейно независимыми собственными векторами, то она должна иметь по крайней мере $(r - 1)$ нулевых поддиагональных элементов (см. гл. 7, § 41). Тем не менее H может иметь собственные значения, совпадающие с рабочей точностью, и при этом каждый из поддиагональных элементов может и не быть мал. Линейно независимые собственные векторы H могут быть получены в этом случае выполнением обратных итераций со слегка различными величинами p , и рассуждения будут те же, что и в § 59 гл. 5.

Если H имеет нелинейный элементарный делитель $(\lambda - \lambda_1)^r$, то векторы, полученные при использовании различных значений p , близких к λ_1 , будут сами близки. Обратные итерации не показывают характерную «нестабильность», связанную с повторяющимся линейным делителем, которая позволяла нам получать линейно независимые собственные векторы. Обычно, если H имеет квадратичный делитель, вычисленные собственные значения не равны, а отличаются на величину порядка $2^{-1/2} \|H\|$ (см. гл. 2, § 23). Собственные векторы x_1 и x_2 , определенные обратными итерациями с использованием этих двух различных собственных значений, отличаются на величины порядка $2^{-1/2}$, и $(x_1 - x_2)$ дает корневой вектор высоты два, но, естественно, он имеет в лучшем случае только $\frac{1}{2}t$ правильных знаков (см. гл. 3, § 63).

Обратные итерации с ленточными матрицами

56. Обратные итерации еще более экономичны в случае трехдиагональных матриц. Мы уже обсуждали этот вопрос в главе 5 в связи с симметричными матрицами, но так как симметрия не дает никаких преимуществ, то комментарии немедленно применимы к несимметричному случаю, за исключением того, что теперь некоторые собственные значения могут быть комплексными, даже если матрица вещественная.

Как мы видели в главе 6, нельзя без дополнительных предположений гарантировать, что обычные методы приведения матриц к трехдиагональному виду будут устойчивыми. Поэтому, если трехдиагональная матрица получена из матрицы общего вида при помощи промежуточной матрицы Хессенберга, мы рекомендуем находить собственные векторы x

матрицы Хессенберга, а не трехдиагональной. Более того, если собственный вектор матрицы Хессенберга уже вычислен, его можно использовать для улучшения собственного значения при помощи прямых итераций, так как оно могло быть определено сравнительно неточно, если переход от формы Хессенберга к трехдиагональной был сравнительно неустойчив. Мы подчеркиваем, что если мы нашли собственный вектор трехдиагональной матрицы, то в любом случае, для того чтобы получить собственный вектор исходной матрицы, мы должны определить по нему собственный вектор матрицы Хессенберга. Поэтому не будет неэкономно определять собственный вектор матрицы Хессенберга непосредственно.

В задачах на собственные значения, связанных с уравнениями в частных производных, часто возникают ленточные матрицы ширины более трех. Обычно такие матрицы симметричны и положительно определены, но так как мы работаем с $(A - pI)$, положительная определенность обязательно утрачивается, если только p не есть приближение к наименьшему собственному значению. Поэтому при разложении в произведение двух треугольных должны быть использованы перестановки, так что получается разложение в произведение треугольных матрицы $P(A - pI)$, где P — матрица перестановок. Если

$$P(A - pI) = LU, \quad (56.1)$$

то легко проверить, что когда A является ленточной матрицей ширины $(2m + 1)$, матрица L имеет лишь m ненулевых поддиагональных элементов в каждом столбце (за исключением, конечно, последних m столбцов), а U имеет не более $(2m + 1)$ ненулевых элементов в каждой строке. Соблазнительно, но неразумно опустить перестановки, особенно если нужно определить несколько средних собственных значений. Заметим, что эти замечания справедливы независимо от того, симметрична или нет матрица A .

Этот метод был использован на ACE для нахождения нескольких собственных векторов симметричной ленточной матрицы порядка 1500 и ширины 31, начиная с довольно хороших собственных значений. Использование отношения Релея дало возможность получить собственные значения с очень высокой точностью при очень малом количестве дополнительной работы.

Комплексно сопряженные собственные векторы

57. Вычисление комплексных собственных векторов вещественных матриц особенно интересно с точки зрения анализа ошибок. Предположим, что $(\xi_1 + i\eta_1)$ — комплексное собственное значение, x_1 — соответствующий ему комплексный собственный вектор, и мы имеем приближение $\lambda = \xi + i\eta$ к $(\xi_1 + i\eta_1)$. Обратные итерации могут быть проведены несколькими способами.

(i) Мы можем вычислить последовательность комплексных векторов z_s и w_s вида

$$(A - \xi - i\eta)w_{s+1} = z_s, \quad z_{s+1} = w_{s+1}/\max(w_{s+1}), \quad (57.1)$$

где все вычисления проведены с использованием комплексных чисел. Вообще говоря, w_s стремятся к комплексному собственному вектору.

(ii) Мы можем работать лишь с вещественной арифметикой следующим образом. Положим $w_{s+1} = u_{s+1} + iw_{s+1}$, $z_s = p_s + iq_s$ и приравняем

вещественные и мнимые части в первом из уравнений (57.1). Это дает

$$(A - \xi I) u_{s+1} + \eta v_{s+1} = p_s, \quad (57.2)$$

$$-\eta u_{s+1} + (A - \xi I) v_{s+1} = q_s, \quad (57.3)$$

что можно записать в виде

$$C(\lambda) \begin{bmatrix} u_{s+1} \\ v_{s+1} \end{bmatrix} = \begin{bmatrix} A - \xi I & \eta I \\ -\eta I & A - \xi I \end{bmatrix} \begin{bmatrix} u_{s+1} \\ v_{s+1} \end{bmatrix} = \begin{bmatrix} p_s \\ q_s \end{bmatrix}. \quad (57.4)$$

Мы получили систему вещественных уравнений с $2n$ переменными.

(iii) Из уравнений (57.2) и (57.3) мы можем вывести уравнения

$$[(A - \xi I)^2 + \eta^2 I] u_{s+1} = (A - \xi I) p_s - \eta q_s, \quad (57.5)$$

$$[(A - \xi I)^2 + \eta^2 I] v_{s+1} = \eta p_s + (A - \xi I) q_s. \quad (57.6)$$

Матрица в обеих этих системах уравнений одинакова; она вещественная и имеет порядок n . Следовательно, требуется лишь одно разложение вещественной матрицы порядка n в произведение двух треугольных.

(iv) Векторы u_{s+1} и v_{s+1} могут быть определены или решением (57.5) с использованием (57.2), или решением (57.6) с использованием (57.3).

Если $(\xi + i\eta)$ не является точным собственным значением, то ни одна из матриц, встречающихся в этих четырех способах, не будет особенной. Каждое из решений единственно, и все они идентичны. Если они на практике отличаются, то это результат ошибок округления. Замечательно, что методы (i), (ii) и (iv) дают сходимость к комплексному собственному вектору, а (iii) — нет!

58. Прежде чем провести анализ ошибок, определим собственные векторы, соответствующие различным уравнениям.

Матрица $(A - \lambda I)$, как матрица в поле комплексных чисел, имеет собственные значения $(\lambda_1 - \lambda)$ и $(\bar{\lambda}_1 - \lambda)$ и соответствующие собственные векторы x_1 и $\bar{x}_1$. Если λ стремится к λ_1 , то первое собственное значение стремится к нулю, а второе — нет. Следовательно, ранг $(A - \lambda_1 I)$, вообще говоря, равен $(n - 1)$.

Матрица $B(\lambda) = (A - \xi I)^2 + \eta^2 I$ — полином от A и, следовательно, также имеет собственные векторы x_1 и $\bar{x}_1$, соответствующие собственным значениям $(\lambda_1 - \xi)^2 + \eta^2$ и $(\bar{\lambda}_1 - \xi)^2 + \eta^2$. Если λ стремится к λ_1 , то оба эти собственных значения стремятся к нулю, и следовательно, вообще говоря, ранг $B(\lambda_1)$ равен $(n - 2)$ и она имеет два линейно независимых собственных вектора x_1 и $\bar{x}_1$.

Возвращаясь к матрице $C(\lambda)$ из (57.4), мы непосредственно проверяем, что

$$\begin{bmatrix} A - \xi I & \eta I \\ -\eta I & A - \xi I \end{bmatrix} \begin{bmatrix} x_1 \\ \pm ix_1 \end{bmatrix} = [(\lambda_1 - \xi) \pm i\eta] \begin{bmatrix} x_1 \\ \pm ix_1 \end{bmatrix}, \quad (58.1)$$

$$\begin{bmatrix} A - \xi I & \eta I \\ -\eta I & A - \xi I \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \pm i\bar{x}_1 \end{bmatrix} = [(\bar{\lambda}_1 - \xi) \pm i\eta] \begin{bmatrix} \bar{x}_1 \\ \pm i\bar{x}_1 \end{bmatrix}. \quad (58.2)$$

Следовательно, $C(\lambda)$ имеет четыре собственных значения $(\lambda_1 - \xi) \pm i\eta$ и $(\bar{\lambda}_1 - \xi) \pm i\eta$ и четыре соответствующих линейно независимых

собственных вектора. Имеем

$$\left. \begin{aligned} (\lambda_1 - \xi) - i\eta &\rightarrow 0 \\ (\bar{\lambda}_1 - \bar{\xi}) + i\eta &\rightarrow 0 \end{aligned} \right\} \text{ при } \xi + i\eta \rightarrow \lambda_1, \quad (58.3)$$

что показывает, что $C(\lambda_1)$ имеет два нулевых собственных значения, которым соответствуют два линейно независимых собственных вектора $[x_1^T; -ix_1^T]$ и $[\bar{x}_1^T; i\bar{x}_1^T]$, и, следовательно, ранг $C(\lambda_1)$ равен $(2n - 2)$.

Анализ ошибок

59. Эти простые результаты позволяют нам весьма просто увидеть, почему именно метод (iii) не имеет успеха на практике. Когда мы выполняем обратные итерации с произвольной матрицей, имеющей два собственных значения, очень близких к нулю, компоненты по всем векторам, за исключением компонент по векторам, соответствующим этим собственным значениям, быстро уменьшаются. Поэтому итерированные векторы лежат в подпространстве, порожденном этими собственными векторами. Конкретный вектор, получающийся в этом подпространстве, сильно зависит от ошибок округления и даже может меняться от итерации к итерации.

Метод (i) не обладает никакими интересными особенностями. Матрица имеет лишь одно малое собственное значение, и мы имеем быструю сходимость к x_1 . Применяется комплексная арифметика; необходимо приблизительно $\frac{1}{3}n^3$ комплексных умножений при разложении матрицы в произведение треугольных и затем n^2 комплексных умножений при каждой итерации.

В методе (ii) вектор $[u_s^T; v_s^T]$, получаемый на каждом этапе, необходимо вещественный. Он сходится к вектору из подпространства, порожденного $[x_1^T; -ix_1^T]$ и $[\bar{x}_1^T; i\bar{x}_1^T]$, и следовательно, должен иметь вид

$$\alpha [x_1^T; -ix_1^T] + \bar{\alpha} [\bar{x}_1^T; i\bar{x}_1^T].$$

Это дает для комплексного собственного вектора

$$(\alpha x_1 + \bar{\alpha} \bar{x}_1) + i(-i\alpha x_1 + i\bar{\alpha} \bar{x}_1) = 2\alpha x_1, \quad (59.1)$$

что вполне удовлетворительно.

В методе (iii) векторы u_s и v_s , очевидно, будут вещественными и будут сходиться к вектору из подпространства, натянутого на x_1 и $\bar{x}_1$. Следовательно, когда компоненты по остальным векторам уже исчезнут (что произойдет почти сразу, если λ близко к λ_1), то

$$u_s = \alpha x_1 + \bar{\alpha} \bar{x}_1, \quad v_s = \beta x_1 + \bar{\beta} \bar{x}_1, \quad (59.2)$$

но мы не можем ожидать, что α и β как-либо связаны. Комплексный собственный вектор, определяемый этим методом, имеет вид

$$u_s + iv_s = (\alpha + i\beta)x_1 + (\bar{\alpha} + i\bar{\beta})\bar{x}_1 \quad (59.3)$$

и он не равен кратному x_1 . Хотя u_s и v_s суть вещественные векторы, порождающие то же подпространство, что и x_1 и $\bar{x}_1$, вектор $(u_s + iv_s)$ не кратен x_1 , и для получения последнего требуются дополнительные вычисления.

В методе (iv) мы сразу после u_{s+1} будем получать вещественный вектор v_{s+1} , связанный с u_{s+1} соотношением (57.2). В предельном случае,

когда λ равно λ_1 с рабочей точностью, мы найдем, что решение u_{s+1} уравнения (57.5) больше, чем p_s и q_s , по которым оно определяется, на множитель порядка $1/\|F\|$, где F — матрица ошибок, возникших при разложении $(A - \xi I)^2 + \eta^2 I$ в произведение треугольных. Следовательно, можно пренебречь p_s в уравнении (57.2), и вычисленный v_{s+1} в основном удовлетворяет соотношению

$$\begin{aligned} \eta v_{s+1} &= -(A - \xi I) u_{s+1} = -(A - \xi I)(\alpha x_1 + \bar{\alpha} \bar{x}_1) = \\ &= \alpha (\xi - \lambda_1) x_1 + \bar{\alpha} (\xi - \bar{\lambda}_1) \bar{x}_1. \end{aligned} \quad (59.4)$$

Тогда (58.3) дает

$$\eta(u_{s+1} + iv_{s+1}) = \alpha \eta x_1 + \bar{\alpha} \eta \bar{x}_1 + \alpha \eta x_1 - \bar{\alpha} \eta \bar{x}_1 = 2\alpha \eta x_1, \quad (59.5)$$

и это показывает, что $(u_{s+1} + iv_{s+1})$ действительно кратен x_1 .

Тот факт, что (iii) дает неправильно связанные u_{s+1} и v_{s+1} , особенно очевиден в случае матрицы второго порядка. В самом деле, легко видеть, что если даже $(p_s + iq_s)$ — точный собственный вектор, $(u_{s+1} + iv_{s+1})$ будет совершенно неудовлетворительным. Это следует из того, что $B(\lambda_1) = 0$ при $n = 2$. Следовательно, если мы вычислим $B(\lambda)$ для значения λ , верного с рабочей точностью, то $B(\lambda)$ будет состоять лишь из ошибок округления. Поэтому ее можно записать в виде $2^{-t}P$, где P — произвольная матрица того же порядка величины, что и A^2 . Если $(p_s + iq_s)$ — собственный вектор, то в правой части (57.5) и (57.6) будут $-2\eta_1 q_s$ и $2\eta_1 p_s$ соответственно. Тогда с точностью до постоянного множителя имеем

$$u_{s+1} = -P^{-1}q_s, \quad v_{s+1} = P^{-1}p_s, \quad (59.6)$$

и так как P произвольна, мы не можем ожидать, что u_{s+1} и v_{s+1} будут правильно связаны.

Мы обсудили предельный случай, когда λ верно с рабочей точностью. Если $\|A\| = 1$ и $|\lambda - \lambda_1| \approx 2^{-k}$, то даже в наиболее благоприятных обстоятельствах метод (iii) будет давать собственный вектор в виде $(x_1 + \epsilon \bar{x}_1)$, где ϵ порядка 2^{k-t} . При k , равных t , компонента по $\bar{x}_1$ становится того же порядка, что и по x_1 .

Численный пример

60. В табл. 3 мы показываем приложение методов (ii), (iii) и (iv) § 57 для вычисления комплексного собственного вектора вещественной матрицы третьего порядка. Обратные итерации выполнялись с $\lambda = 0,82197 + 0,71237i$, что совпадает с точным значением с рабочей точностью. Сначала даем разложение в произведение треугольных с перестановками вещественной матрицы $C(\lambda)$ шестого порядка из (57.4); треугольные матрицы обозначены L_1 и U_1 . Тот факт, что $C(\lambda)$ имеет два собственных значения, близких к нулю, проявляется наиболее очевидным образом в том, что последние два диагональных элемента U_1 равны нулю. При первой итерации можем начать с обратного хода и в качестве правой части взять e . Ясно, что последние две компоненты решения могут быть произвольно большими и с любым отношением. Соответствующие компоненты в правой части фактически игнорируются. Мы вычислили два решения, выбрав последние две компоненты равными 0 и 10^6 в одном случае, и 10^6 и 10^6 — в другом. Хотя вещественные решения получились существенно различными, соответствующие нормированные комплексные собственные векторы оба равны истинному собственному вектору с рабочей точностью.

Таблица 3

$$A = \begin{bmatrix} 0,59797 & 0,29210 & 0,15280 \\ -0,95537 & 0,75756 & -1,23999 \\ 0,50608 & 0,13706 & 0,40117 \end{bmatrix} \quad \lambda = 0,82197 + 0,71237 i$$

 L_1

$$\begin{bmatrix} 1,00000 & & & & & \\ 0,00000 & 1,00000 & & & & \\ -0,52972 & -0,14450 & 1,00000 & & & \\ 0,00000 & 0,00000 & 0,66104 & 1,00000 & & \\ 0,23446 & -0,43124 & -0,41157 & 0,40774 & 1,00000 & \\ 0,74565 & -0,06742 & -0,85798 & -0,68110 & 0,00000 & 1,00000 \end{bmatrix}$$

 U_1

$$\begin{bmatrix} -0,95537 & -0,06441 & -1,23999 & 0,00000 & 0,71237 & 0,00000 \\ & -0,71237 & 0,00000 & -0,95537 & -0,06441 & -1,23999 \\ & & -1,07765 & -0,13805 & 0,36805 & 0,53319 \\ & & & 0,59734 & -0,10624 & -0,77326 \\ & & & & 0,00000 & 0,00001 \\ & & & & & 0,00000 \end{bmatrix}$$

(1-е решение) $\times 10^{-6}$

-0,19254
-3,47674
0,32894
.....
1,29451
0,00000
1,00000

(2-е решение) $\times 10^{-6}$

0,16158
-3,80567
0,64769
.....
1,47286
1,00000
1,00000

Нормированные 1-е и 2-е решения

0,05538 - 0,37233 i
1,00000 + 0,00000 i
-0,09461 - 0,28763 i

 $(A - \xi I)^2 + \eta^2 I$

$$\begin{bmatrix} 0,35591 & -0,06330 & -0,46073 \\ -0,35200 & 0,06260 & 0,45568 \\ -0,45726 & 0,08132 & 0,59192 \end{bmatrix}$$

 b_1

$$\begin{bmatrix} -0,49147 \\ -2,97214 \\ -0,49003 \end{bmatrix}$$

 b_2

$$\begin{bmatrix} 0,93327 \\ -1,54740 \\ 0,93471 \end{bmatrix}$$

 L_2

$$\begin{bmatrix} 1,00000 & & \\ -1,28476 & 1,00000 & \\ -0,98901 & 0,00000 & 1,00000 \end{bmatrix}$$

 U_2

$$\begin{bmatrix} 0,35591 & -0,06330 & -0,46073 \\ -0,00001 & -0,00001 & \\ & 0,00001 & \end{bmatrix}$$

Решение 1-е из уравнений (57.5) и (57.2)

-366219 - 228762 i
457966 - 1051690 i
-345821 - 32222 i

Нормированное 1-е решение

0,05538 - 0,37234 i
1,00000 + 0,00000 i
-0,09461 - 0,28763 i

Решение 2-е из уравнений (57.6) и (57.3)

-41426 - 107670 i
266728 - 150935 i
-68649 - 62439 i

Нормированное 2-е решение

0,05540 - 0,37232 i
1,00000 + 0,00000 i
-0,09461 - 0,28763 i

Решение 3-е из уравнений (57.5) и (57.6)

-366219 - 107670 i
457966 - 150935 i
-345821 - 62439 i

Нормированное 3-е решение не аппроксимирует собственный вектор

Затем даем разложение в произведение двух треугольных вещественной матрицы $B(\lambda)$ третьего порядка; треугольные матрицы обозначены L_2 и U_2 . Снова тот факт, что $B(\lambda)$ имеет очень близкие к нулю два собственных значения, проявляется в появлении двух, почти равных нулю диагональных элементов у U_2 . Мы выполнили один шаг, взяв $(e + ei)$ в качестве начального комплексного вектора. В методе (iii) правые части $(A - \xi I)e - \eta e$ и $(A - \xi I)e + \eta e$ обозначены b_1 и b_2 . Оба вектора, полученные по методу (iv), верны с точностью до двух единиц последнего знака. Вектор, полученный по методу (iii), совершенно неточный; он является линейной комбинацией x_1 и x_1 .

Обобщенная проблема собственных значений

61. Метод обратных итераций немедленно распространяется на обобщенную проблему собственных значений (гл. 1, § 30). Достаточно рассмотреть наиболее частый случай определения λ и x таких, что

$$(A\lambda^2 + B\lambda + C)x = 0. \quad (61.1)$$

Так как этот случай эквивалентен стандартной проблеме собственных значений

$$\left[\begin{array}{c|c} 0 & I \\ \hline -A^{-1}C & -A^{-1}B \end{array} \right] \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}, \quad (61.2)$$

соответствующая итерационная процедура имеет вид

$$\left[\begin{array}{c|c} -\lambda I & I \\ \hline -A^{-1}C & -A^{-1}B - \lambda I \end{array} \right] \begin{bmatrix} x_{s+1} \\ y_{s+1} \end{bmatrix} = \begin{bmatrix} x_s \\ y_s \end{bmatrix}, \quad (61.3)$$

где λ — приближенное собственное значение (для удобства мы опустили нормировку). Из уравнения (61.3) имеем

$$\left. \begin{aligned} -\lambda x_{s+1} + y_{s+1} &= x_s, \\ -A^{-1}Cx_{s+1} - A^{-1}By_{s+1} - \lambda y_{s+1} &= y_s, \end{aligned} \right\} \quad (61.4)$$

и, следовательно,

$$-C(y_{s+1} - x_s) - \lambda By_{s+1} - \lambda^2 Ay_{s+1} = \lambda Ay_s, \quad (61.5)$$

$$(\lambda^2 A + \lambda B + C)y_{s+1} = Cx_s - \lambda Ay_s. \quad (61.6)$$

Векторы y_{s+1} и x_{s+1} могут быть вычислены из уравнений (61.6) и (61.4) соответственно. Требуется лишь одно разложение в произведение двух треугольных для матрицы $(\lambda^2 A + \lambda B + C)$.

Этот процесс исключительно эффективен, когда A , B и C являются функциями параметра V и мы хотим проследить зависимость собственного значения $\lambda_1(V)$ и соответствующего ему собственного вектора $x_1(V)$ от V . Величины $\lambda_1(V)$ и $x_1(V)$ для предыдущего значения V могут быть использованы в качестве хорошего начального приближения к их текущим значениям.

Изменение приближенных собственных значений

62. До сих пор мы обсуждали обратные итерации, используя фиксированные приближения к каждому собственному значению, так что при определении каждого собственного вектора употреблялось лишь одно разложение в произведение треугольных. Это весьма эффективная процедура, если приближение к собственному значению хорошее. Скорость сходимости при этом такова, что вряд ли стоит использовать преимущества имеющегося после выполнения итерации более хорошего приближения.

Существует другой подход, при котором обратные итерации используются для первоначальной локализации собственного значения. Если A — симметричная, соответствующий итерационный процесс определяется

(см. гл. 3, § 54) уравнениями

$$\left. \begin{aligned} (A - \mu_s I) v_{s+1} &= u_s, & \mu_{s+1} &= v_{s+1}^T A v_{s+1} / v_{s+1}^T v_{s+1}, \\ u_{s+1} &= v_{s+1} / (v_{s+1}^T v_{s+1})^{1/2}. \end{aligned} \right\} \quad (62.1)$$

На каждой стадии дается новое приближение μ_{s+1} к собственному значению отношением Релея, соответствующим текущему вектору. Легко показать, что этот процесс кубически сходится в окрестности каждого собственного вектора. Предположим, мы имеем

$$u_s = x_k + \sum' \epsilon_i x_i, \quad (62.2)$$

где $\sum'$ означает, что индекс k пропущен при суммировании. Тогда

$$\mu_s = (\lambda_k + \sum' \lambda_i \epsilon_i^2) / (1 + \sum' \epsilon_i^2), \quad (62.3)$$

и, следовательно,

$$v_{s+1} = x_k / (\lambda_k - \mu_s) + \sum' \epsilon_i x_i / (\lambda_i - \mu_s). \quad (62.4)$$

С точностью до нормирующего множителя имеем

$$\begin{aligned} u_{s+1} &= x_k + (\lambda_k - \mu_s) \sum' \epsilon_i x_i / (\lambda_i - \mu_s) = \\ &= x_k - \left\{ \frac{\sum' (\lambda_i - \lambda_k) \epsilon_i^2}{1 + \sum' \epsilon_i^2} \right\} \sum' \epsilon_i x_i / (\lambda_i - \mu_s), \end{aligned} \quad (62.5)$$

и все коэффициенты, за исключением коэффициента при x_k , кубические по ϵ_i .

Существует аналогичный процесс для несимметричных матриц, при котором на каждой стадии определяются приближенные правый и левый собственные векторы. Процесс имеет вид

$$\left. \begin{aligned} (A - \mu_s I) v_{s+1} &= u_s, & (A - \mu_s I)^T q_{s+1} &= p_s, \\ \mu_{s+1} &= q_{s+1}^T A v_{s+1} / q_{s+1}^T v_{s+1}, \\ u_{s+1} &= v_{s+1} / \max(v_{s+1}), & p_{s+1} &= q_{s+1} / \max(q_{s+1}). \end{aligned} \right\} \quad (62.6)$$

Разложение матрицы $(A - \mu_s I)$ в произведение треугольных дает возможность определить v_{s+1} и q_{s+1} . Текущее значение μ_{s+1} есть теперь обобщенное отношение Релея, соответствующее q_{s+1} и v_{s+1} (гл. 3, § 58).

Процесс снова сходится кубически. Предположим, что мы имеем

$$u_s = x_k + \sum' \epsilon_i x_i, \quad p_s = y_k + \sum' \eta_i y_i. \quad (62.7)$$

Тогда

$$\mu_s = \frac{(\lambda_k y_k^T x_k + \sum' \lambda_i \epsilon_i \eta_i y_i^T x_i)}{(y_k^T x_k + \sum' \epsilon_i \eta_i y_i^T x_i)}. \quad (62.8)$$

и, следовательно, с точностью до нормирующих множителей

$$u_{s+1} = x_k + (\lambda_k - \mu_s) \sum' \epsilon_i x_i / (\lambda_i - \mu_s), \quad (62.9)$$

$$p_{s+1} = y_k + (\lambda_k - \mu_s) \sum' \eta_i y_i / (\lambda_i - \mu_s). \quad (62.10)$$

Все коэффициенты, кроме коэффициентов при x_k и y_k , кубичны по ϵ_i и η_i . Если число обусловленности $y_k^T x_k$ собственного значения λ_k мало, то начало кубической сходимости задерживается.

Скорость сходимости, очевидно, весьма хорошая, хотя довольно часто область кубической сходимости достигается тогда, когда количество разря-

дов, с которым проводятся вычисления уже ограничивает возможное увеличение точности. Если требуются точные собственные значения, они могут быть получены аккуратным вычислением последнего отношения Релея или обобщенного отношения Релея. Обычно отношение Релея, полученное на этой стадии, совпадает с точным лучше, чем с одинарной точностью. Отношение может быть найдено точно без использования двойной точности, если мы вычислим величины

$$(A - \mu_s I) u_{s+1}, \quad \mu_{s+1} = \mu_s + u_{s+1}^T (A - \mu_s I) u_{s+1} / u_{s+1}^T u_{s+1}, \quad (62.11)$$

или в несимметричном случае

$$(A - \mu_s I) u_{s+1}, \quad \mu_{s+1} = \mu_s + p_{s+1}^T (A - \mu_s I) u_{s+1} / p_{s+1}^T u_{s+1}. \quad (62.12)$$

Вектор $(A - \mu_s I) u_{s+1}$ должен быть вычислен с использованием накопления скалярных произведений. Обычно происходит взаимное уничтожение, и $(A - \mu_s I) u_{s+1}$ одинарной точности может быть использован при определении μ_{s+1} .

Островский (1958, 1959) дал весьма детальный анализ этих методов в серии статей, к которым читатель отсылается за дальнейшей информацией.

Определение и левых, и правых собственных векторов требует много времени. Если известно достаточно хорошее приближение μ_s к собственному значению, то сходимость обычного процесса обратных итераций во всяком случае будет более чем удовлетворительной. Прямые итерации, обратные итерации и исчерпывание могут использоваться в комбинации следующим образом. Прямые итерации используются до тех пор, пока не будет получено достаточно хорошее значение μ_s и некоторое приближение для собственного вектора. Затем применяются обратные итерации с $(A - \mu_s I)$, начиная с последнего вектора, полученного при прямых итерациях. Окончательно после определения собственного вектора производится процесс исчерпывания. Эта техника особенно эффективна, если A — комплексная матрица. Великолепное описание ее использования содержится у Осборна (1958).

Уточнение системы собственных значений

63. В заключение мы рассмотрим уточнение вычисленной системы собственных значений. В § 51 мы показали, что обычно можем определить собственный вектор, соответствующий ближайшему к наперед заданному числу p собственному значению, с любой желаемой точностью при помощи обратных итераций. Хотя на практике этот метод дает убедительные доказательства точности вычисленного собственного вектора, его невозможно использовать для получения точных оценок.

В гл. 3, §§ 59—64 мы обсуждали проблему уточнения полной системы собственных значений и собственных векторов с теоретической точки зрения. Сейчас опишем детально практическую процедуру, использующую описанные там идеи. Эта процедура была запрограммирована на АСЕ с использованием фиксированной запятой, и мы будем описывать ее в этой форме. Изменения, необходимые для вычислений с плавающей запятой, очевидны.

Предположим, что каким-то образом мы определили приближенную систему собственных значений и собственных векторов. Обозначим вычисленные собственные значения и собственные векторы через λ'_i и x'_i , а матрицу собственных векторов через X' , и определим матрицу невязки F

соотношением

$$AX' = X' \text{diag}(\lambda_i) + F, \quad (X')^{-1}AX' = \text{diag}(\lambda_i) + (X')^{-1}F. \quad (63.1)$$

Предполагаем, что $\|F\| \ll \|A\|$, иначе вряд ли можно рассматривать λ_i и x_i как приближенные собственные значения и собственные векторы. На АСЕ каждый столбец f_i матрицы F вычислялся *точно* с накоплением скалярных произведений. Элементы f_i имели двойное количество знаков, хотя, по нашему предположению относительно F , большинство цифр в первой половине знаков каждого элемента было нулями. Поэтому каждый вектор f_i выражался в форме $2^{-k_i} g_i$, где k_i выбран так, что максимальный элемент $|g_i|$ заключается между $1/2$ и 1 . Каждый f_i преобразован в нормализованный блочно-плавающий вектор с двойной точностью.

До сих пор не появилось никаких ошибок округления. Далее, мы хотим вычислить $(X')^{-1}F$ так точно, как это возможно. На АСЕ это производилось с использованием итерационного метода, описанного в гл. 4, § 69. Матрица X' сначала разлагается в произведение треугольных с использованием перестановок, а затем каждое уравнение $X'h_i = g_i$ решается с помощью итераций. Заметим, что метод, описанный в главе 4, использует одинарную точность вычислений с накоплением скалярных произведений, но, как мы указывали, можно использовать правые части с двойной точностью. В предположении, что $n2^{-t} \|(X')^{-1}\| < 1/2$, процесс сходится и дает решения $\bar{h}_i$ уравнений $X'h_i = g_i$, верные с рабочей точностью. Если положим $p_i = 2^{-k_i} h_i$, то будем иметь

$$(X')^{-1}AX' = \text{diag}(\lambda_i) + P + Q = B, \quad (63.2)$$

где каждый столбец p_i матрицы P известен и представлен в виде блочно-плавающего вектора с одинарной точностью. Q точно неизвестна, но

$$\|q_i\|_\infty \leq 2^{-t-1} \|p_i\|_\infty \quad (63.3)$$

из предположения, что итерационная процедура дает правильно округленное решение. Собственные значения A в *точности* совпадают с собственными значениями B .

64. Теперь мы можем для локализации собственных значений использовать теорему Гершгорина (гл. 2, § 13). Так как B масштабирована по столбцам, то на АСЕ удобнее применять обычные результаты Гершгорина к B^T , а не к B , однако мы использовали здесь обычные круги, так что можно сделать прямую ссылку на предыдущее. Непосредственное приложение теоремы Гершгорина показывает, что собственные значения A лежат в кругах с центрами $\lambda'_i + p_{ii} + q_{ii}$ и радиусами $\sum_{j \neq i} |p_{ij} + q_{ij}|$, но, вообще говоря, это очень слабый результат. На АСЕ более точная локализация получалась следующим образом.

Предположим, что круг, связанный с λ'_i , изолирован. Умножим первый столбец B на 2^{+k} и первую строку на 2^{-k} , где k — наибольшее целое число такое, что первый круг Гершгорина остается изолированным. В силу наших предположений $k \geq 0$. При таком значении k теорема Гершгорина утверждает, что одно собственное значение находится в круге с центром $\lambda'_1 + p_{11} + q_{11}$ и радиусом $2^{-k} \sum_{j \neq 1} |p_{1j} + q_{1j}|$, а значит, и в круге с центром $\lambda'_1 + p_{11}$ и радиусом $|q_{11}| + 2^{-k} \sum_{j \neq 1} (|p_{1j}| + |q_{1j}|)$. Мы можем заменить $|q_{1j}|$ их верхней границей, так как в любом случае они меньше, чем $|p_{1j}|$, в 2^t раз.

Следующие соображения позволяют выбрать подходящее k весьма просто. Радиус первого круга Гершгорина равен приблизительно $2^{-k} \sum_{j \neq 1} |p_{1j}|$, в то время как главные части остальных радиусов равны приблизительно $2^k |p_{ii}|$, если k заметно больше единицы. Поэтому на АСЕ k сначала выбирается наибольшим целым числом таким, что

$$2^k |p_{ii}| < |\lambda_i - \lambda_j| \quad (i = 2, \dots, n), \quad (64.1)$$

а это очень простое требование. Затем проверяем, будет ли при этом первый круг изолирован. Это почти всегда так, но если это нарушится, то k нужно уменьшить так, чтобы первый круг был изолирован. Аналогичный процесс может быть использован для локализации каждого собственного значения в предположении, что соответствующий круг Гершгорина матрицы B изолирован.

Численный пример

65. Прежде чем описывать аналогичный процесс для собственных векторов, проиллюстрируем технику определения собственных значений на простом примере. В табл. 4 приближенная система, состоящая из λ_i

Таблица 4

$\begin{bmatrix} -0,05811 & -0,11695 & 0,51004 & -0,31330 \\ -0,04291 & 0,56850 & -0,07041 & -0,68747 \\ -0,01652 & 0,38953 & 0,01203 & -0,52927 \\ 0,06140 & 0,32179 & -0,22094 & -0,42448 \end{bmatrix}$				
x'_1	x'_2	x'_3	x'_4	
0,38463	0,59641	1,00000	1,00000	
0,91066	1,00000	0,74534	0,34879	
0,67339	0,81956	0,46333	0,52268	
1,00000	0,23761	0,61755	0,10392	
λ'	$10^5 f_1$	$10^5 f_2$	$10^5 f_3$	$10^5 f_4$
-0,25660	-0,06427	0,32258	0,09052	-0,14193
0,32185	-0,22972	0,30806	-0,17442	0,13411
-0,10244	-0,19421	0,13829	-0,10132	0,05223
0,13513	-0,12232	0,05163	-0,09736	0,15437
L				
$\begin{bmatrix} 1,00000 & & & \\ 0,91066 & 1,00000 & & \\ 0,38463 & 0,64447 & 1,00000 & \\ 0,67339 & 0,81168 & -0,16525 & 1,00000 \end{bmatrix}$				
Перестановка: строки 4, 2, 1, 3 переставляются на место 1, 2, 3, 4				
U				
$\begin{bmatrix} 1,00000 & 0,23761 & 0,61755 & 0,10392 \\ & 0,78362 & 0,18296 & 0,25415 \\ & & 0,64456 & 0,79624 \\ & & & 0,37037 \end{bmatrix}$				
$10^5 P$ ($P = \text{округленное } (X')^{-1} F$)				
$\begin{bmatrix} -0,14241 & -0,31415 & -0,11093 & 0,18613 \\ -0,17215 & 0,29352 & -0,18228 & 0,07080 \\ 0,09990 & 0,52207 & 0,06180 & -0,04284 \\ -0,00672 & -0,25371 & 0,18010 & -0,21291 \end{bmatrix}$				

Элементы $|q_{ij}|$ не превосходят $\frac{1}{2} 10^{-10}$

Продолжение табл. 4

Корни	Возможный множитель	Собственное значение	Граница ошибки
1	10^{-5}	-0,25660 14241	$10^{-10} (0,61121) + 2 \times 10^{-15}$
2	10^{-4}	0,32185 29352	$10^{-9} (0,42523) + 2 \times 10^{-14}$
3	10^{-5}	-0,10243 93820	$10^{-10} (0,66481) + 2 \times 10^{-15}$
4	10^{-5}	0,13512 78709	$10^{-10} (0,44053) + 2 \times 10^{-15}$
Первый собственный вектор $\Lambda' + (X')^{-1}F$		Первый собственный вектор A	
1,00000		0,38462 67268 $\pm 12 \times 10^{-10}$	
$10^{-5} (0,29760) \pm 4 \times 10^{-10}$		0,91066 11902 $\pm 15 \times 10^{-10}$	
$10^{-5} (-0,64801) \pm 7 \times 10^{-10}$		0,67339 17329 $\pm 12 \times 10^{-10}$	
$10^{-5} (0,01716) \pm 3 \times 10^{-10}$		1,00000 00000	

и x'_i , дает в точности выписанные векторы-невязки f_i . Они выражены в блочно-плавающей десятичной форме. Матрица X' была разложена в произведение треугольных с использованием перестановок. При этом получились матрицы L и U , и перестановки строк были записаны. Матрица P является правильно округленной $(X')^{-1} F$. Как часто случается, один и тот же масштабный множитель подходит для всех столбцов P . Элементы неизвестной матрицы Q ограничены $\frac{1}{2} 10^{-10}$. Для локализации каждого собственного значения был использован множитель вида 10^{-k} . Условие (64.1) дает оптимальное значение для 10^{-k} в каждом случае, и в таблице выписаны соответствующие границы для исправленных собственных значений. Все вычисления выполнены с использованием пятизначной десятичной арифметики с накоплением там, где это необходимо, скалярных произведений, и при этом оказалось возможным дать границы ошибок, которые оказались порядка 10^{-10} .

Уточнение собственных векторов

66. Если собственные векторы B равны u_i , то собственные векторы A суть $X'u_i$. Следовательно, для получения собственных векторов A достаточно определить собственные векторы B . Вычисление этих собственных векторов и определение границы для ошибок лишь немногим более сложно, но так как запись значительно более громоздкая, мы ограничимся показом этой техники на численном примере. Предположим, что мы имеем

$$B = \begin{bmatrix} 0,93256 & & \\ & 0,61321 & \\ & & 0,11763 \end{bmatrix} + 10^{-5} \begin{bmatrix} 0,31257 & 0,41321 & 0,21653 \\ 0,41357 & 0,21631 & -0,40257 \\ 0,31765 & 0,41263 & 0,71654 \end{bmatrix} + Q,$$

$$|q_{ik}| < \frac{1}{2} 10^{-10}. \quad (66.1)$$

С помощью техники для собственных значений мы можем показать, что первое собственное значение μ удовлетворяет условию

$$|\mu - 0,9325631257| \leq 10^{-9}. \quad (66.2)$$

Пусть u — соответствующий собственный вектор, нормированный так, что его элемент, наибольший по модулю, равен единице. Эта компонента

обязательно должна быть первой, так как если мы предположим, что это вторая (или третья), то второе (или третье) уравнение $Bu = \mu u$ немедленно приведет к противоречию.

Если мы положим $u^T = (1, u_2, u_3)$, то второе и третье уравнения дают

$$10^{-5}(0,41375) + q_{21} + (0,6132121631 + q_{22})u_2 + \\ + [-10^{-5}(0,40257) + q_{23}]u_3 = \mu u_2, \quad (66.3)$$

$$10^{-5}(0,31765) + q_{31} + [10^{-5}(0,41263) + q_{32}]u_2 + \\ + [0,1176371654 + q_{33}]u_3 = \mu u_3. \quad (66.4)$$

Так как $|u_2|, |u_3| \leq 1$, эти уравнения вместе с (66.2) дают

$$|u_2| < \frac{10^{-5}(0,41375) + \frac{1}{2}10^{-10} + 10^{-5}(0,40257) + \frac{1}{2}10^{-10}}{\lambda'_1 - \lambda'_2 - \frac{1}{2}10^{-10} - 10^{-9}} < 3 \cdot 10^{-5}, \quad (66.5)$$

причем мы использовали весьма грубые оценки. Аналогично

$$|u_3| < 2 \cdot 10^{-5}. \quad (66.6)$$

Используя эти грубые оценки, мы можем вернуться к уравнениям (66.3), (66.4) и получить более близкое приближение к u_2 и u_3 . Будет проще, если мы перепишем (66.3) в виде

$$\left(10^{-5}a \pm \frac{1}{2}10^{-10}\right) + \left(\lambda'_2 \pm \frac{1}{2}10^{-10}\right)u_2 + \left(10^{-5}c \pm \frac{1}{2}10^{-10}\right)u_3 = \\ = (\lambda'_1 \pm 10^{-9})u_2, \quad (66.7)$$

где a и c порядка единицы, и элементы $\pm 10^{-k}$ показывают неопределенность в коэффициентах. Из (66.7) имеем

$$u_2 = \frac{\left(10^{-5}a \pm \frac{1}{2}10^{-10}\right) + \left(10^{-5}c \pm \frac{1}{2}10^{-10}\right)u_3}{\lambda'_1 - \lambda'_2 \pm 10,5 \cdot 10^{-10}}, \quad (66.8)$$

и мы уже знаем, что u_3 лежит между $\pm 2 \cdot 10^{-5}$. Взяв наиболее неблагоприятную комбинацию знаков, получаем

$$u_2 < \frac{10^{-5}a + \frac{1}{2}10^{-10} + 2 \cdot 10^{-5} \left(-10^{-5}c + \frac{1}{2}10^{-10}\right)}{\lambda'_1 - \lambda'_2 - 10,5 \cdot 10^{-10}} < \\ < \frac{10^{-5}(0,41377)}{0,31935} < 10^{-5}(1,2957), \quad (66.9)$$

$$u_2 > \frac{10^{-5}a - \frac{1}{2}10^{-10} - 2 \cdot 10^{-5} \left(-10^{-5}c + \frac{1}{2}10^{-10}\right)}{\lambda'_1 - \lambda'_2 + 10,5 \cdot 10^{-10}} > \\ > \frac{10^{-5}(0,41373)}{0,31936} > 10^{-5}(1,2954). \quad (66.10)$$

Аналогично

$$10^{-5}(0,38982) > u_3 > 10^{-5}(0,38976). \quad (66.11)$$

Следовательно, мы получили границы ошибок, которые меньше $2 \cdot 10^{-9}$ для каждой компоненты первого собственного вектора.

Вычисление собственных векторов на настольной вычислительной машине не особенно длинно, но оно настолько неприятно, что мы вычислили только первый собственный вектор. Мы даем первый собственный вектор и для B и для A , с границами ошибок для каждого. Границы для ошибок в компонентах векторов не больше, чем $1^{1/2}$ единицы девятого десятичного знака, и можно ожидать, что они существенно завышены. На самом деле максимальная ошибка в любой компоненте не превосходит двух единиц десятого знака. Но весьма трудно получить такие точные оценки без усложненного анализа ошибок. На вычислительной машине типа АСЕ, которая имеет 48 двоичных разрядов, потеря нескольких двоичных разрядов из 96 сравнительно несущественна. Предельные оценки, как правило, значительно более точные, чем мы обычно требуем.

Комплексно сопряженные собственные значения

67. Рассмотрения предыдущего параграфа очевидным образом могут быть распространены на комплексные матрицы. В этом случае все вычисления производятся в поле комплексных чисел, но никаких существенных новых черт не появляется. Если мы имеем вещественную матрицу с одной или более комплексно сопряженной парой собственных значений, то естественно, что мы хотим работать в поле вещественных чисел как можно больше. Соответствующая техника хорошо иллюстрируется примером матрицы четвертого порядка с одной парой комплексно сопряженных собственных значений. Тогда мы имеем

$$AX' = X' \begin{bmatrix} \lambda' & \mu' & & \\ -\mu' & \lambda' & & \\ & & \lambda'_3 & \\ & & & \lambda'_4 \end{bmatrix} + F, \quad (67.1)$$

где первые два столбца X' суть вещественная и мнимая части вычисленного комплексного собственного вектора. Как и раньше, мы можем вычислить такую матрицу P , что

$$(X')^{-1}AX' = \begin{bmatrix} \lambda' & \mu' & & \\ -\mu' & \lambda' & & \\ & & \lambda'_3 & \\ & & & \lambda'_4 \end{bmatrix} + P + Q, \quad (67.2)$$

причем P известна точно, и мы имеем оценки для элементов Q . До сих пор все элементы вещественны. Если

$$D = \begin{bmatrix} 1 & 1 & \vdots & \\ i & -i & & \\ \hdashline & & 1 & \\ & & & 1 \end{bmatrix}, \quad D^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -i & \vdots & \\ 1 & i & & \\ \hdashline & & 2 & \\ & & & 2 \end{bmatrix}, \quad (67.3)$$

то

$$D^{-1}(X')^{-1}AX'D = \begin{bmatrix} \lambda' + i\mu' & & & \\ & \lambda' - i\mu' & & \\ & & \lambda'_3 & \\ & & & \lambda'_4 \end{bmatrix} + D^{-1}PD + D^{-1}QD \quad (67.4)$$

и

$$D^{-1}PD = \frac{1}{2} \begin{bmatrix} p_{11} + p_{22} + i(p_{12} - p_{21}) & p_{11} - p_{22} - i(p_{12} + p_{21}) & p_{13} - ip_{23} & p_{14} - ip_{24} \\ p_{11} - p_{22} + i(p_{12} + p_{21}) & p_{11} + p_{22} - i(p_{12} - p_{21}) & p_{13} + ip_{23} & p_{14} + ip_{24} \\ \hline p_{31} + ip_{32} & p_{31} - ip_{32} & 2p_{33} & 2p_{34} \\ p_{41} + ip_{42} & p_{41} - ip_{42} & 2p_{43} & 2p_{44} \end{bmatrix}. \quad (67.5)$$

Центрами двух соответствующих кругов Гершгорина будут

$$\lambda' + \frac{1}{2}(p_{11} + p_{22}) \pm i \left[\mu' + \frac{1}{2}(p_{12} - p_{21}) \right],$$

и мы имеем исправленные собственные значения. Мы можем уменьшить радиус первого круга, как прежде, умножая первую строку на 2^{-k} и первый столбец на 2^k . По аналогии с (64.1) выберем теперь k так, чтобы

$$\left. \begin{aligned} \frac{1}{2} 2^k |p_{11} - p_{22} + i(p_{12} + p_{21})| &< 2|\mu'|, \\ \frac{1}{2} 2^k |p_{j1} + ip_{j2}| &< |\lambda' - \lambda_j + i\mu'| \quad (j > 2). \end{aligned} \right\} \quad (67.6)$$

Совпадающие и патологически близкие собственные значения

68. Описанные методы дают малые границы для ошибок для любого собственного значения, которое достаточно хорошо отделено от остальных. Если разделение двух или более собственных значений ухудшается, то и оценки ошибок тоже ухудшаются. Эти оценки могут быть исправлены следующим образом.

Рассмотрим матрицу $(\Lambda' + \varepsilon B)$, где Λ' — диагональная и Λ' и B имеют элементы порядка единицы. Предположим, что первые два диагональных элемента λ'_1 и λ'_2 матрицы Λ' отделены величиной порядка ε . Представим $(\Lambda' + \varepsilon B)$ в виде

$$\begin{bmatrix} \lambda'_1 & & \\ & \lambda'_2 & \\ \hline & & \Lambda'_{n-2} \end{bmatrix} + \varepsilon \begin{bmatrix} B_{22} & B_{2, n-2} \\ \hline B_{n-2, 2} & B_{n-2, n-2} \end{bmatrix}. \quad (68.1)$$

Если

$$C_{22} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} + (\lambda'_2 - \lambda'_1)\varepsilon^{-1} \end{bmatrix}, \quad (68.2)$$

то (68.1) можно записать в виде

$$\begin{bmatrix} \lambda'_1 & & \\ & \lambda'_1 & \\ \hline & & \Lambda'_{n-2} \end{bmatrix} + \varepsilon \begin{bmatrix} C_{22} & B_{2, n-2} \\ \hline B_{n-2, 2} & B_{n-2, n-2} \end{bmatrix}, \quad (68.3)$$

где C_{22} имеет элементы порядка единицы. Предположим теперь, что мы вычислили систему собственных значений и собственных векторов для C_{22} , так что

$$C_{22}X_{22} = X_{22} \begin{bmatrix} \mu_1 & \\ & \mu_2 \end{bmatrix} + \varepsilon G_{22}, \quad (68.4)$$

где G_{22} порядка единицы. Тогда имеем

$$X_{22}^{-1}C_{22}X_{22} = \begin{bmatrix} \mu_1 & \\ & \mu_2 \end{bmatrix} + \varepsilon H_{22}, \quad \text{где} \quad H_{22} = X_{22}^{-1}G_{22}. \quad (68.5)$$

Снова можно ожидать, что элементы H_{22} будут порядка единицы, если только проблема *собственных значений* для C_{22} не слишком плохо обусловлена. Следовательно, положив

$$Y = \left[\begin{array}{c|c} X_{22} & O \\ \hline O & I \end{array} \right], \quad (68.6)$$

имеем

$$Y^{-1}(\Lambda' + \varepsilon B)Y = \begin{bmatrix} \lambda_1 + \varepsilon \mu_1 & & \\ & \lambda_1' + \varepsilon \mu_2 & \\ \hline & & \Lambda_{n-2} \end{bmatrix} + \\ + \varepsilon \begin{bmatrix} \varepsilon H_{22} & X_{22}^{-1}B_{2, n-2} \\ \hline B_{n-2, 2}X_{22} & B_{n-2, n-2} \end{bmatrix}. \quad (68.7)$$

Вторая матрица имеет элементы порядка ε^2 в верхнем левом углу и элементы порядка ε в остальных позициях.

Рассмотрим теперь специальный пример. Предположим, что матрица (68.7) имеет вид

$$\begin{bmatrix} 0,92563 \ 41653 & & \\ & 0,92563 \ 21756 & \\ & & 0,11325 \end{bmatrix} + \\ + 10^{-5} \begin{bmatrix} 10^{-5}(0,23127) & 10^{-5}(0,31257) & 0,31632 \\ 10^{-5}(0,41365) & 10^{-5}(0,41257) & 0,21753 \\ \hline 0,28316 & 0,10175 & 0,20375 \end{bmatrix}. \quad (68.8)$$

Если мы умножим первые две строки на 10^{-5} , а первые два столбца — на 10^5 , то круги Гершгорина будут:

центр 0,92563 41653 23127, радиус $10^{-10}(0,31257 + 0,31632)$;

центр 0,92563 21756 41257, радиус $10^{-10}(0,41365 + 0,21753)$;

центр 0,11325 20375, радиус $(0,28316 + 0,10175)$.

Все три круга сейчас разделены, и первые два собственных значения даны с ошибкой порядка 10^{-10} . В этом примере мы опустили матрицу ошибок, соответствующую Q ; в соответствии с предыдущими рассмотрениями она имела бы элементы порядка 10^{-10} . Это дало бы вклад порядка 10^{-10} в наши границы для ошибок.

Вообще говоря, если r собственных значений патологически близки, то мы способны разделить их, решая независимую проблему собственных значений порядка r , используя ту же точность вычислений, что и в основной проблеме. Если два или более собственных значений на самом деле совпадают, то соответствующие круги будут продолжать пересекаться, но границы ошибок этим методом будут улучшаться.

Замечания о программах на АСЕ

69. На основе приемов, описанных в предыдущих параграфах, были разработаны несколько программ на АСЕ. Они были разной степени сложности. Наиболее сложные из них давали более точное разделение патологически близких собственных значений, как описано в § 68.

Для большинства матриц, даже для таких, у которых разделение собственных значений нормально могло бы рассматриваться как очень плохое, границы ошибок были значительно меньше, чем обычно требуется. Например, для матрицы 15-го порядка, имеющей элементы порядка единицы, с собственными значениями, разделенными величинами 10^{-8} , 10^{-7} , 10^{-7} , 10^{-6} , 10^{-5} , 10^{-5} , 10^{-5} , 10^{-5} , 10^{-2} , 10^{-2} , 10^{-1} , 1, все собственные значения и собственные векторы гарантируются с точностью до 17 десятичных знаков, причем с значительно лучшей гарантией для большинства из них. Заметим, что основным ограничением для всех программ было требование, чтобы была возможность вычислять $(X')^{-1}F$ с рабочей точностью. Когда число обусловленности X' по отношению к обращению достигает $2^t/n$, это становится все более трудным, и когда это значение достигается, процесс нарушается и становится существенной более высокая точность вычислений. Такая неприятность на АСЕ встречалась только для матриц, для которых проблема собственных значений очень плохо обусловлена.

Программу можно было использовать итерационно, так что после вычисления исправленной системы собственных значений все величины округлялись до одинарной точности и процесс применялся снова. При использовании такого способа сложные вычисления, необходимые для получения границ ошибок, опускались на всех шагах, кроме последнего, и процесс можно начинать со сравнительно плохих приближений. Эта итерационная техника имеет много общего с методами Магнье (1948), Яна (1948) и Коллара (1948).

В этих методах матрица P , определенная из равенства

$$(X')^{-1}AX' = \text{diag}(\lambda'_i) + P = B, \quad (69.1)$$

вычисляется без каких-либо специальных предосторожностей. Исправленные собственные значения A берутся равными $(\lambda'_i + p_{ii})$, а i -й собственный вектор B (гл. 2, § 6) равным

$$[p_{1i}/(\lambda'_i - \lambda'_1); p_{2i}/(\lambda'_i - \lambda'_2); \dots; 1; \dots; p_{ni}/(\lambda'_i - \lambda'_n)]. \quad (69.2)$$

Исправленные собственные векторы A вычисляются обычным способом.

Независимо методы предыдущих параграфов были разработаны (Уилкинсон, 1961а) главным образом с целью определения строгих границ для вычисляемых величин. Надо подчеркнуть, что когда приближенная система собственных значений и собственных векторов определена одним

из более точных методов, которые мы обсуждали, ошибки обычно того же порядка, что и при прямом применении уравнений (69.1) и (69.2). При получении исправленной системы собственных значений и собственных векторов существенно более внимательное отношение к деталям, воплощенным в программах на АСЕ.

Дополнительные замечания

Использование итерационных методов, в которых одновременно участвует несколько векторов, было описано в работах Бауэра (1957 и 1958), и я следовал в основном его исследованиям. Метод § 38 был позднее исследован Воеводиным (1962). Большая часть материала этой главы также обсуждалась в главе 7 книги Хаусхолдера (1964) с несколько отличной точки зрения.

Ступенчатые итерации, метод ортогонализации и биортогонализации могут быть осуществлены с использованием обратной матрицы. В частности, в случае ступенчатых итераций мы имеем

$$AX_{s+1} = L_s, \quad X_{s+1} = L_{s+1}R_{s+1},$$

и, очевидно, не нужно определять A^{-1} точно. Нужно лишь осуществить разложение A в произведение треугольных. Это замечание особенно важно, если A в форме Хессенберга. Мы можем применить сдвиги, но при этом требуется разложение $(A - pI)$ в произведение треугольных для каждого нового значения p . При этом возникает трудность, отмеченная в § 34, что p не может свободно выбираться после того, как λ_i уже определено.

ЛИТЕРАТУРА

- Арнольд (Arnoldi W. E.), 1951, The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.* 9, 17—29.
- Бауэр (Bauer F. L.), 1957, Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. *Z. Angew. Math. Phys.* 8, 214—235.
- Бауэр (Bauer F. L.), 1958, On modern matrix iteration processes of Bernoulli and Graeffe type. *J. Ass. Comp. Math.* 5, 246—257.
- Бауэр (Bauer F. L.), 1959, Sequential reduction to tridiagonal form. *J. Soc. Industr. Appl. Math.* 7, 107—113.
- Бауэр (Bauer F. L.), 1963, Optimally scaled matrices. *Numerische Math.* 5, 73—87.
- Бауэр и Файк (Bauer F. L. and Fike C. T.), 1960, Norms and exclusion theorems. *Numerische Math.* 2, 137—141.
- Бауэр и Хаусхолдер (Bauer F. L. and Householder A. S.), 1960, Moments and characteristic roots. *Numerische Math.* 2, 42—53.
- Беллман (Bellman R.), 1960, Introduction to Matrix Analysis. McGraw-Hill, New York.
- Берстоу (Bairstow L.), 1914, The solution of algebraic equations with numerical coefficients. *Rep. Memor. Adv. Comm. Aero., Lond.*, No. 154, 51—64.
- Бирхгоф и Маклейн (Birkhoff G. and MacLane S.), 1953, A Survey of modern algebra (revised edition). Macmillan, New York.
- Бодевиг (Bodewig E.), 1959, Matrix Calculus (second edition). Interscience Publishers, New York; North-Holland Publishing Company, Amsterdam.
- Брукер и Сумнер (Brooker R. A. and Sumner F. H.), 1956, The method of Lanczos for calculating the characteristic roots and vectors of a real symmetric matrix. *Proc. Instn. Elect. Engrs B103 Suppl.* 114—119.
- Варга (Varga R. S.), 1962, Matrix Iterative Analysis. Prentice-Hall, New Jersey.
- Виландт (Wielandt H.), 1944a, Bestimmung höherer Eigenwerte durch gebrochene Iteration. *Ber. B44/J/37 der Aerodynamischen Versuchsanstalt Göttingen.*
- Виландт (Wielandt H.), 1944b, Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben. *Math. Z.* 50, 93—143.
- Воеводин В. В., 1962, Некоторые методы решения полной проблемы собственных значений, *ЖВМ и МФ* 2, 15—24.
- Гантмахер Ф. Р., 1967, Теория матриц, М., «Наука».
- Гастинель (Gastinel N.), 1960, Matrices du second degré et normes générales en analyse numérique linéaire. Thesis, University of Grenoble.
- Гершгорин (Gerschgorin S.), 1931, Über die Abgrenzung der Eigenwerte einer Matrix. *Изв. АН СССР, сер. физ.-матем.* 6, 749—754.
- Гивенс (Givens W.), 1953, A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices. *Appl. Math. Ser. nat. Bur. Stand.* 29, 117—122.
- Гивенс (Givens W.), 1954, Numerical computation of the characteristic values of a real symmetric matrix. Oak Ridge National Laboratory, ORNL-1574.
- Гивенс (Givens W.), 1957, The characteristic value-vector problem. *J. Ass. Comp. Mach.* 4, 298—307.
- Гивенс (Givens W.), 1958, Computation of plane unitary rotations transforming a general matrix to triangular form. *J. Soc. Industr. Appl. Math.* 6, 26—50.
- Гивенс (Givens W.), 1959, The linear equations problem. *Applied Mathematics and Statistics Laboratories, Stanford University, Tech. Rep. No. 3.*
- Голдстейн и Хорвиц (Goldstine H. H. and Horwitz L. P.), 1959, A procedure for the diagonalization of normal matrices. *J. Ass. Comp. Mach.* 6, 176—195.
- Голдстейн, Мюррей и Нейман (Goldstine H. H., Murray F. J. and von Neumann J.), 1959, The Jacobi method for real symmetric matrices. *J. Ass. Comp. Mach.* 6, 59—96.

- Голдстейн и Нейман (Goldstine H. H. and von Neumann J.), 1951, Numerical inverting of matrices of high order, II. Proc. Amer. Math. Soc. 2, 188—202.
- Гофман и Виландт (Hoffman A. J. and Wielandt H. W.), 1953, The variation of the spectrum of a normal matrix, Duke Math. J. 20, 37—39.
- Грегори (Gregory R. T.), 1953, Computing eigenevalues and eigenvectors of a symmetric matrix on the ILLIAC. Math. Tab., Wash. 7, 215—220.
- Грегори (Gregory R. T.), 1958, Results using Lanczos' method for finding eigenevalues of arbitrary matrices. J. Soc. Industr. Appl. Math. 6, 182—188.
- Гринштадт (Greenstadt J.), 1955, A method for finding roots of arbitrary matrices. Math. Tab., Wash. 9, 47—52.
- Гринштадт (Greenstadt J.), 1960, The determination of the characteristic roots of a matrix by the Jacobi method. Mathematical Methods for Digital Computers (edited by Ralston and Wilf), 84—91. Wiley, New York.
- Гурса (Goursat E.), 1933, Cours d'Analyse Mathématique, Tome II (fifth edition). Gauthier Villars, Paris. (Русский перевод: Гурса Э., 1936, Курс математического анализа, том 2, изд. 3-е, М.—Л., Глав. ред. общетехн. лит. номографии.)
- Данилевский А., 1937, О численном решении векового уравнения. Матем. сб. 2(44), 169—171.
- Деккер (Dekker T. J.), 1962, Series AP 200 Algorithms. Mathematisch Centrum, Amsterdam.
- Дин (Dean P.), 1960, Vibrational spectra of diatomic chains. Proc. Roy. Soc. A. 254, 507—521.
- Дюран (Durand E.), 1960, Solutions Numériques des Equations Algébriques. T. 1. Masson, Paris.
- Дюран (Durand E.), 1961, Solutions Numériques des Equations Algébriques. T. 2. Masson, Paris.
- Иогансен (Johansen D. E.), 1961, A modified Givens method for the eigenvalue evaluation of large matrices. J. Ass. Comp. Mach. 8, 331—335.
- Като (Kato T.), 1949, On the upper and lower bounds of eigenevalues. J. Phys. Soc. Japan 4, 334—339.
- Кинкайд (Kincaid W. M.), 1947, Numerical methods for finding characteristic roots and vectors of matrices. Quart. Appl. Math. 5, 320—345.
- Козей (Causey R. L.), 1958, Computing eigenevalues of non-Hermitian matrices by methods of Jacobi type. J. Soc. Industr. Appl. Math. 6, 172—181.
- Козей и Грегори (Causey R. L. and Gregory R. T.), 1961, On Lanczos' algorithm for tridiagonalizing matrices. SIAM Rev. 3, 322—328.
- Коллар (Collar A. R.), 1948, Some notes on Jahn's method for the improvement of approximate latent roots and vectors of a square matrix. Quart. J. Mech. 1, 145—148.
- Коллатц (Collatz L.), 1948, Eigenwertprobleme und Ihre Numerische Behandlung. Chelsea, New York.
- Крылов А. Н., 1931, О численном решении уравнения, которым в технических вопросах определяются частоты малых колебаний материальных систем, Изв. Акад. наук СССР, сер. физ.-мат., 4, 491—539.
- Кублановская В. Н., 1961, О некоторых алгоритмах для решения полной проблемы собственных значений, ЖВМ и МФ 1, 555—570.
- Курант и Гильберт (Courant R. and Hilbert D.), 1953, Methods of Mathematical Physics, vol. 1. Interscience Publishers, New York. (Русский перевод нем. издания: Курант Р., Гильберт Д., 1951, Методы математической физики, т. 1, М.—Л., Гостехиздат.)
- Ланцоз (Lanczos C.), 1950, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Stand. 45, 255—282.
- Леверье (Le Verrier U. J. J.), 1840, Sur les variations seculaires des éléments elliptiques des sept planètes principales. Liouville J. Maths 5, 220—254.
- Лоткин (Lotkin M.), 1956, Characteristic values of arbitrary matrices. Quart. Appl. Math. 14, 267—275.
- Магнье (Magnier A.), 1948, Sur le calcul numérique des matrices. C. R. Acad. Sci., Paris, 226, 464—465.
- Маели (Machly H. J.), 1954, Zur iterativen Auflösung algebraischer Gleichungen. Z. Angew. Math., Phys. 5, 260—263.
- Маккиман (McKeeman W. M.), 1962, Crout with equilibration and iteration. Commun. Ass. Comp. Math. 5, 553—555.

- Маркус (Marcus M.), 1960, Basic theorems in matrix theory. Appl. Math. Ser. nat. Bur. Stand. 57.
- Мюллер (Muller D. E.), 1956, A method for solving algebraic equations using an automatic computer. Math. Tab. Wash., 10, 208—215.
- Нейман и Голдстейн (von Neumann J. and Goldstine H. H.), 1947, Numerical inverting of matrices of high order. Bull. Amer. Math. Soc. 53, 1021—1099.
- Ньюман (Newman M.), 1962, Matrix computations. A Survey of numerical analysis (edited by Todd), 222—254. McGraw-Hill, New York.
- Олвер (Oliver F. W. J.), 1952, The evaluation of zeros of high-degree polynomials. Phil. Trans. Roy. Soc. A244, 385—415.
- Ортега (Ortega J. M.), 1963, An error analysis of Householder's method for the symmetric eigenvalue problem. Numerische Math. 5, 211—225.
- Ортега и Райзер (Ortega J. M. and Kaiser H. F.), 1963, The LLT and QR -methods for symmetric tridiagonal matrices. Computer J. 6, 99—101.
- Осборн (Osborne E. E.), 1958, On acceleration and matrix deflation processes used with the power method. J. Soc. Industr. Appl. Math. 6, 279—287.
- Осборн (Osborne E. E.), 1960, On pre-conditioning of matrices. J. Ass. Comp. Mach. 7, 338—345.
- Островский (Ostrowski A. M.), 1957, Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen. Jber. Deutsch. Math. Verein. 60, Abt. 1, 40—42.
- Островский (Ostrowski A. M.), 1958—1959, On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors. I, II, III, IV, V, VI, Arch. rat. Mech. Anal. 1, 233—241; 2, 423—428; 3, 325—340; 3, 341—347; 3, 472—481; 4, 153—165.
- Островский (Ostrowski A. M.), 1960, Solution of equations and systems of equations. Academic Press, New York and London. (Русский перевод: Островский А. М., 1963, Решение уравнений и систем уравнений, ИЛ).
- Парлетт (Parlett B.), 1964, Laguerre's method applied to the matrix eigenvalue problem. Maths Computation 18, 464—485.
- Попе и Томпкинс (Pope D. A. and Tompkins C.), 1957, Maximizing functions of rotations-experiments concerning speed of diagonalization of symmetric matrices using Jacobis method. J. Ass. Comp. Mach. 4, 459—466.
- Ричардсон (Richardson L. F.), 1950, A purification method for computing the latent columns of numerical matrices and some integrals of differential equations. Phil. Trans. Roy. Soc. A242, 439—491.
- Роллет и Уилкинсон (Rollett J. S. and Wilkinson J. H.), 1961, An efficient scheme for the codiagonalization of a symmetric matrix by Givens'-method in a computer with a two-level store. Computer J. 4, 177—180.
- Россер, Ланцос, Хестенес и Каруш (Rosser B. J., Lanczos C., Hestenes M. R. and Karush W.), 1951, The separation of close eigenvalues of a real symmetric matrix. J. Res. Nat. Bur. Stand. 47, 291—297.
- Рутисхаузер (Rutishauser H.), 1953, Beiträge zur Kenntnis des Biorthogonalisierungs-Algorithmus von Lanczos, Z. angew. Math. Phys. 4, 35—56.
- Рутисхаузер (Rutishauser H.), 1955, Bestimmung der Eigenwerte und Eigenvektoren einer Matrix mit Hilfe des Quotienten-Differenzen-Algorithmus. Z. angew. Math. Phys. 6, 387—401.
- Рутисхаузер (Rutishauser H.), 1957, Der Quotienten-Differenzen-Algorithmus. Birkhäuser, Basel/Stuttgart. (Русский перевод: Рутисхаузер Г., 1960, Алгоритмы частных и разностей, ИЛ).
- Рутисхаузер (Rutishauser H.), 1958, Solution of eigenvalue problems with the LR -transformation. Appl. Math. Ser. nat. Bur. Stand. 49, 47—81.
- Рутисхаузер (Rutishauser H.), 1959, Deflation bei Bandmatrizen. Z. angew. Math. Phys. 10, 314—319.
- Рутисхаузер (Rutishauser H.), 1960, Über eine kubisch konvergente Variante der LR -Transformation. Z. angew. Math. Mech. 40, 49—54.
- Рутисхаузер (Rutishauser H.), 1963, On Jacobi rotation patterns. Proceedings A. M. S. Symposium in Applied Mathematics 15, 219—239.
- Рутисхаузер и Шварц (Rutishauser H. and Schwarz H. R.), 1963, Handbook Series Linear Algebra. The LR -transformation method for symmetric matrices. Numerische Math. 5, 273—289.
- Страчей и Френсис (Strachey C. and Francis J. G. F.), 1961, The reduction of a matrix to codiagonal form by eliminations. Computer J. 4, 168—176.

- Тарнове (Tarnove I.), 1958, Determination of eigenvalues of matrices having polynomial elements. *J. Soc. Industr. Appl. Math.* **6**, 163—171.
- Таусский (Taussky O.), 1949, A recurring theorem on determinants. *Amer. Math. Mon.* **56**, 672—676.
- Таусский (Taussky O.), 1954, Contributions to the solution of systems of linear equations and the determination of eigenvalues. *Appl. Math. Ser. nat. Bur. Stand.* **39**.
- Таусский и Маркус (Taussky O. and Marcus M.), 1962, Eigenvalues of finite matrices. *A Survey of Numerical Analysis* (edited by Todd), 279—313. McGraw-Hill, New York.
- Темпле (Temple G.), 1952, The accuracy of Rayleigh's method of calculating the natural frequencies of vibrating systems. *Proc. Roy. Soc.* **A211**, 204—224.
- Тодд (Todd J.), 1949, The condition of certain matrices, *1. Quart. J. Mech.* **2**, 469—472.
- Тодд (Todd J.), 1950, The condition of certain matrices. *Proc. Camb. Phil. Soc.* **46**, 116—118.
- Тодд (Todd J.), 1954a, The condition of certain matrices. *Arch. Math.* **5**, 249—257.
- Тодд (Todd J.), 1954b, The condition of the finite segments of the Hilbert matrix. *Appl. Ser. Math. nat. Bur. Stand.* **39**, 109—116.
- Тодд (Todd J.), 1961, Computational problems concerning the Hilbert matrix. *J. Res. Nat. Bur. Stand.* **65**, 19—22.
- Турнбулл (Turnbull H. W.), 1950, *The Theory of Determinants, Matrices, and Invariants*. Blackie, London and Glasgow.
- Турнбулл и Эйткин (Turnbull H. W. and Aitken, A. C.), 1932, *An Introduction to the Theory of Canonical Matrices*. Blackie, London and Glasgow.
- Трауб (Traub J. F.), 1964, *Iterative Methods for the Solution of Equations*. Prentice-Hall, New Jersey.
- Тьюринг (Turing A. M.), 1948, Rounding-off errors in matrix processes. *Quart. J. Mech.* **1**, 287—308.
- Уайт (White P. A.), 1958, The computation of eigenvalues and eigenvectors of a matrix. *J. Soc. Industr. Appl. Math.* **6**, 393—437.
- Уилкинсон (Wilkinson J. H.), 1954, The calculation of the latent roots and vectors of matrices on the pilot model of the ACE. *Proc. Camb. Phil. Soc.* **50**, 536—566.
- Уилкинсон (Wilkinson J. H.), 1958a, The calculation of the eigenvectors of codiagonal matrices. *Computer J.* **1**, 90—96.
- Уилкинсон (Wilkinson J. H.), 1958b, The calculation of eigenvectors by the method of Lanczos. *Computer J.* **1**, 148—152.
- Уилкинсон (Wilkinson J. H.), 1959a, The evaluation of the zeros of ill-conditioned polynomials, Parts I and II. *Numerische Math.* **1**, 150—180.
- Уилкинсон (Wilkinson J. H.), 1959b, Stability of the reduction of a matrix to almost triangular and triangular forms by elementary similarity transformations. *J. Ass. Comp. Mach.* **6**, 336—359.
- Уилкинсон (Wilkinson J. H.), 1960a, Householder's method for the solution of the algebraic eigenproblem. *Computer J.* **3**, 23—27.
- Уилкинсон (Wilkinson J. H.), 1960b, Error analysis of floating-point computation. *Numerische Math.* **2**, 319—340.
- Уилкинсон (Wilkinson J. H.), 1960c, Rounding errors in algebraic processes. *Information Processing*, 44—53. Unesco, Paris, Butterworths, London.
- Уилкинсон (Wilkinson J. H.), 1961a, Rigorous error bounds for computed eigensystems. *Computer J.* **4**, 230—241.
- Уилкинсон (Wilkinson J. H.), 1961b, Error analysis of direct methods of matrix inversion. *J. Ass. Comp. Mach.* **8**, 281—330.
- Уилкинсон (Wilkinson J. H.), 1962a, Instability of the elimination method of reducing a matrix to tri-diagonal form. *Computer J.* **5**, 61—70.
- Уилкинсон (Wilkinson J. H.), 1962b, Error analysis of eigenvalue techniques based on orthogonal transformations. *J. Soc. Industr. Appl. Math.* **10**, 162—195.
- Уилкинсон (Wilkinson J. H.), 1962c, Note on the quadratic convergence of the cyclic Jacobi process. *Numerische Math.* **4**, 296—300.
- Уилкинсон (Wilkinson J. H.), 1963a, Plane-rotations in floating-point arithmetic. *Proceedings A.M.S. Symposium in Applied Mathematics* **15**, 185—198.
- Уилкинсон (Wilkinson J. H.), 1963b, Rounding Errors in Algebraic Processes, Notes on Applied Science No. 32, Her Majesty's Stationary Office, London, Prentice-Hall, New Jersey.
- Фаддеев Д. К., Фаддеева В. Н., 1963, *Вычислительные методы линейной алгебры*, Физматгиз, Москва.

- Фаддеева В. Н., 1950, Вычислительные методы линейной алгебры. Гостехиздат.
- Феллер и Форсайт (Feller W. and Forsythe G. E.), 1951, New matrix transformations for obtaining characteristic vectors. *Quart. Appl. Math.* 8, 325—331.
- Фокс (Fox L.), 1954, Practical solution of linear equation and inversion of matrices. *Appl. Math. Ser. nat. Bur. Stand.* 39, 1—54.
- Фокс (Fox L.), 1964, An Introduction to Numerical Linear Algebra. Oxford University Press, London.
- Фокс, Хускей и Уилкинсон (Fox L., Huskey H. D. and Wilkinson J. H.), 1948, Notes on the solution of algebraic linear simultaneous equations. *Quart. J. Mech.* 1, 149—173.
- Форсайт (Forsythe G. E.), 1958, Singularity and near singularity in numerical analysis. *Amer. Math. Mon.* 65, 229—240.
- Форсайт (Forsythe G. E.), 1960, Crout with pivoting. *Commun. Ass. Comp. Mach.* 3, 507—508.
- Форсайт и Вазов (Forsythe G. E. and Wasow W. R.), 1960, Finite-Difference Methods for Partial Differential Equations, Wiley, New York, London. (Русский перевод: Вазов В., Форсайт Дж., 1963, Разностные методы решения дифференциальных уравнений в частных производных, ИЛ.)
- Форсайт и Страус (Forsythe G. E. and Straus E. G.), 1955, On best conditioned matrices. *Proc. Amer. Math. Soc.* 6, 340—345.
- Форсайт и Страус (Forsythe G. E. and Straus L. W.), 1955, The Souriau-Frame characteristic equation algorithm on a digital computer. *J. Maths. Phys.* 34, 152—156.
- Форсайт и Хенричи (Forsythe G. E. and Henrici P.), 1960, The cyclic Jacobi method for computing the principal values of a complex matrix. *Trans. Amer. Math. Soc.* 94, 1—23.
- Франк (Frank W. L.), 1958, Computing eigenvalues of complex matrices by determinant evaluation and by methods of Danilewski and Wielandt. *J. Soc. Industr. Appl. Math.* 6, 378—392.
- Фрезер, Дункан и Коллар (Frazer R. A., Duncan W. J. and Collar A. R.), 1947, Elementary Matrices and Some Applications to Dynamics and Differential Equations. Macmillan, New York. (Русский перевод: Фрезер Р., Дункан В., Коллар А., 1950, Теория матриц и ее приложения к дифференциальным уравнениям и динамике, ИЛ.)
- Френсис (Francis J. G. F.), 1961, 1962, The QR transformation, Parts I and II. *Computer J.* 4, 265—271, 332—345.
- Хандскомб (Handscorn D. C.), 1962, Computation of the latent roots of a Hessenberg matrix by Bairstow's method. *Computer J.* 5, 139—141.
- Хаусхолдер (Householder A. S.), 1958a, The approximate solution of matrix problems. *J. Ass. Comp. Mach.* 5, 205—243.
- Хаусхолдер (Householder A. S.), 1958b, Unitary triangularization of a nonsymmetric matrix. *J. Ass. Comp. Mach.* 5, 339—342.
- Хаусхолдер (Householder A. S.), 1958c, Generated error in rotational tridiagonalization. *J. Ass. Comp. Mach.* 5, 335—338.
- Хаусхолдер (Householder A. S.), 1961, On deflating matrices. *J. Soc. Industr. Appl. Math.* 9, 89—93.
- Хаусхолдер (Householder A. S.), 1964, The Theory of Matrices in Numerical Analysis. Blaisdell, New York.
- Хаусхолдер и Байер (Householder A. S. and Bauer F. L.), 1959, On certain methods for expanding the characteristic polynomial. *Numerische Math.* 1, 29—37.
- Хенричи (Henrici P.), 1958a, On the speed of convergence of cyclic and quasiscyclic Jacobi methods for computing eigenvalues of Hermitian matrices. *J. Soc. Industr. Appl. Math.* 6, 144—162.
- Хенричи (Henrici P.), 1958b, The quotient-difference algorithm. *Appl. Math. Ser. nat. Bur. Stand.* 49, 23—46.
- Хенричи (Henrici P.), 1962, Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Numerische Math.* 4, 24—40.
- Хессенберг (Hessenberg K.), 1941, Auflösung linearer Eigenwertaufgaben mit Hilfe der Hamilton—Cayleyschen Gleichung, Dissertation, Technische Hochschule, Darmstadt.
- Химан (Himan M. A.), 1957, Eigenvalues and eigenvectors of general matrices. Twelfth National meeting A.C.M., Houston, Texas.
- Хотеллинг (Hotelling H.), 1933, Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417—441, 498—520.